

Vanshaj Khattar

☎ +1 540-514-3029 | ✉ vanshajk@vt.edu | 🏠 vanshajkhattar.github.io | 📄 Google Scholar

Summary

Ph.D. candidate with research experience in agentic AI, large language models, and reinforcement learning. Recent work includes building a self-improving agent harness using automated evaluation and execution feedback, supervised fine-tuning and reinforcement learning for large reasoning models. Experienced with Python, PyTorch, Hugging Face Transformers, vLLM, verl, Docker, and SLURM.

Education

Virginia Polytechnic Institute and State University (Virginia Tech)

PH.D. ELECTRICAL ENGINEERING

- Advisor: Dr. Ming Jin
- GPA: 3.64/4.0
- Thesis: Towards Trustworthy Reinforcement Learning Agents.

Blacksburg, VA

August 2021 - November 2026

Virginia Polytechnic Institute and State University (Virginia Tech)

MS ELECTRICAL ENGINEERING

- Advisor: Dr. Azim Eskandarian

Blacksburg, VA

August 2019 - May 2021

Delhi Technological University

B.TECH ELECTRICAL AND ELECTRONICS ENGINEERING

- CGPA: 8.09/10.0

New Delhi, India

August 2014 - May 2018

Industry Experience

Mitsubishi Electric Research Labs (MERL)

RESEARCH SCIENTIST INTERN IN PRIVATE AND SECURE AGENTIC AI

- **Mentors:** Ye Wang, Jing Liu, and Toshiaki Koike-Akino
- **Project 1:** Built a closed-loop self-improving agent system on the OAI harness in the HarnessAudit benchmark: executed tasks, detected and evaluated failures, proposed harness modifications, reevaluated candidate changes, and automatically reverted updates based on performance and safety.
- **Project 2:** Identified reward hacking mechanisms in current test-time steering methods for vision-language-action (VLA) models and proposed mitigation methods using inference-time alignment. **Conference paper under preparation.**

Cambridge, Boston

May 2026 - Present

Mitsubishi Electric Research Labs (MERL)

RESEARCH SCIENTIST INTERN IN TRUSTWORTHY AND GENERAL AI

- **Mentors:** Ye Wang, Jing Liu, and Toshiaki Koike-Akino
- **Project:** Performed supervised fine-tuning and test-time reinforcement learning on large reasoning models; developed jail-break evaluations that exposed safety degradation during post-training; proposed a safe-RL method for mitigation.
- **Achievements:** One workshop paper accepted at **AAAI 2026**, **one patent** and longer version currently under review.

Cambridge, Boston

May 2025 - August 2025

National Renewable Energy Lab (NREL)

GRADUATE SUMMER INTERN IN ML FOR POWER SYSTEMS

- **Mentors:** Yiyun Yao and Fei Ding
- **Project:** Developed a hierarchical multi-agent reinforcement learning framework using graph-based representations for coordinated critical-load restoration under uncertain distribution-grid topology.
- Evaluated agent coordination and decision-making across changing network configurations; work resulted in an IEEE PES-GM 2025 paper and one journal paper under preparation (**Preprint**)

Golden, Colorado

June 2024 - August 2024

Selected Publications

CONFERENCE AND WORKSHOP PUBLICATIONS

Choudhury, M., **Khattar, V.**, Liu, J., Koike-Akino, T., Chakrabarty, A., Zilberstein, S., Wang, Y. "Training Language Models to Cooperate with Inference-Time Controllers". **Preprint under review**

Khattar, V., Choudhury, M., Rashid, M., Liu, J., Koike-Akino, T., Jin, M., Wang, Y., "Amplification Effects in Test-Time Reinforcement Learning: Safety and Reasoning Vulnerabilities". **AAAI 2026-Trustworthy Agentic AI Workshop**

Khattar, V., Yao, Y., Ding, F., Jin, “Distribution Grid Critical Load Restoration under Uncertain Topology Changes via a Hierarchical Multi-Agent Reinforcement Learning Approach”. **IEEE PES-GM 2025**

Sel, B., Al-Tawaha, A., **Khattar, V.**, Jia, R. and Jin, M., “Algorithm of thoughts: Enhancing exploration of ideas in large language models”. (**ICML 2024**)

Khattar, V.*, Lin, T*, Huang. Y*, Jia, R., Hong, J., Liu C, Vincentelli, A and Jin, M., “CausalPrompt: Enhancing LLMs with Weakly Supervised Causal Reasoning for Non-Language Applications”. (**ICLR 2024 Workshop Paper**)

Khattar, V. and Jin, M., “Optimization Solution Functions as Deterministic Policies for Offline Reinforcement Learning”. (American Control Conference) (**ACC 2024**)

Manzoor, F., **Khattar, V.**, Liu, C., and Jin, M., “Zero-day Attack Detection in Digital Substations using In-Context Learning”. (**SmartGridComm 2024**)

Khattar, V., Ding, Y., Sel, B., Lavaei, J. and Jin, M., “A CMDP-within-online framework for Meta-Safe Reinforcement Learning”. In The Eleventh International Conference on Learning Representations (**ICLR 2023 Spotlight**) .

Khattar, V. and Jin, M., “Winning the CityLearn challenge: adaptive optimization with evolutionary search under trajectory-based guidance”. In Proceedings of the (**AAAI 2023**).

Jin, M., **Khattar, V.**, Kaushik, H., Sel, B. and Jia, R., “On solution functions of optimization: universal approximation and covering number bounds”. In Proceedings of the (**AAAI 2023**).

Meimand, M., **Khattar, V.**, Yazdani, Z., Jazizadeh, F., Jin, M., “TUNEOP: An Evolutionary Reinforcement Learning HVAC System Controller For Tuning Energy-Comfort Optimization Formulations”. (**BuildSys 2023**).

Khattar, V. and Eskandarian, A., “Stochastic predictive control for crash avoidance in autonomous vehicles based on stochastic reachable set threat assessment”. (**IMECE 2021**).

Khattar, V. and Eskandarian, A., “Reactive online motion re-planning for crash mitigation in autonomous vehicles using bezier curve optimization”. ASME (**IMECE 2020**).

Valluru, S.K., Singh, M., Singh, M. and **Khattar, V.**, “Experimental validation of PID and LQR control techniques for stabilization of cart inverted pendulum system”. In IEEE International Conference on (**RTEICT 2018**).

Technical Skills

Programming languages. Python, C, MATLAB

AI/ML Frameworks. PyTorch, TensorFlow, vLLM, Scikit-learn, Hugging Face Transformers, OpenAI API

Robotic platforms. MuJoCo, LIBERO, RLBench, OpenVLA

Software engineering and HPC. Git, Linux, SLURM, shell scripting, workflow automation, Docker

Awards & Scholarships

2026	ICML 2026 Silver Reviewer Award, ICML	
2026	Pratt Scholarship Award, ECE Virginia Tech	\$ 950
2023	AAAI 2023 travel scholarship., AAAI	\$ 750
2022	Member of the winning team ROLEVT at CityLearn challenge 2021.(ROLEVT team),	\$ 1500
2021	Second in the Torgersen Research Excellence Award for MS Oral presentation. (Link),	\$ 500

Outreach and Service

Conference reviewer: 1) AISTATS 2022-2026; 2) ICLR 2025, 2026; 3) ICML 2025, 2026; 4) AAAI 2026, 2027

Workshops: Organized Trustworthy Interactive Decision-Making with Foundation Models workshop at IJCAI 2024 ([Link](#))

Tutorials: Safe RL for Smart Grids tutorial at SmartGridComm 2024 conference. ([Link](#))