AAKRITI AGRAWAL

+1 (240) 728-3745
agrawal5@umd.edu sites.google.com/umd.edu/aakriti-agrawal Google Scholar

EDUCATION

University of Maryland, College Park Jan 2021 – Aug 2026
Ph.D. in Computer Science. Advisor: Furong Huang, Dinesh Manocha
Outstanding Achievement Award (2025-26)

Birla Institute of Technology and Science, Pilani Campus May 2019
B.E. in Electrical and Electronics Engineering. GPA: 9.41/10

RESEARCH INTEREST

I am an AI safety researcher focused on building reliable and aligned AI systems, especially for improved reasoning, continual learning, and agentic deployment. I specialize in post-training (RL/SFT), identifying agentic misalignment and safety vulnerabilities, and improving the factuality and groundedness of frontier models.

Aligned and Safe Reasoning in LLMs: identifying hidden bias in process reward models [1] and mitigating reward hacking and downstream policy misalignment in large reasoning models (LRMs) [1,2], safer process supervision for policy learning (GRPO, RLFH, PPO) and policy search [1,2]. I am interested to extend effective process-supervision for better reasoning [1,2,5] and planing as my primary research interest.

Superalignment and Multi-LLM System: improving scalable oversight and weak-to-strong generalization with multiple LLMs [3], multi-LLM reasoning and LLM evaluation using uncertainty-aware answer selection for diverse LLMs [4].

Interpretability and Multimodal Robustness: reducing hallucinations in vision-language models through refined textual embeddings [6], studying diffusion language models for better reasoning [7], and improving robustness in multimodal systems [8,9,11].

I also have background in multi-agent reinforcement learning [11, 12, 13, 14], robotics [13, 14] and speech [15] applications.

SELECTED PUBLICATIONS

Please visit my Google Scholar for the complete list. * denotes equal contribution.

1. **The Hidden Bias of Process Reward Models: PRISM for Rewarding the Right Reasoning** Link
   A. Agrawal, S. Chakraborty, A. Saghafian, N. Sharma, R. Fathony, N. H. Nguyen, C. B. Bruss, A. S. Bedi, F. Huang.
   ICLR 2026 Workshop (AFAA); In review at NeurIPS 2026.
2. **VeriGate: Verifier-Gated Step-Level Supervision for GRPO** Link
   A. Agrawal, M. Liu, F. Huang.
   In review at NeurIPS 2026.
3. **EnsemW2S: Enhancing Weak-to-Strong Generalization with Large Language Model Ensemble** Link
   A. Agrawal, M. Ding, Z. Che, C. Deng, A. Satheesh, B. An, C. B. Bruss, J. Langford, F. Huang.
   ACL 2026; NeurIPS 2024 Workshop (SafeGenAi).
4. **Uncertainty-Aware Answer Selection for Improved Reasoning in Multi-LLM Systems** Link
   A. Agrawal, R. Aralikatti, A. Satheesh, A. Singh Bedi, F. Huang.
   EMNLP 2025.
5. **Easy2Hard-Bench: Standardized Difficulty Labels for Profiling LLM Performance and Generalization** Link
   M. Ding*, C. Deng*, J. Choo, Z. Wu, A. Agrawal, A. Schwarzschild, T. Zhou, T. Goldstein, J. Langford, A. Anandkumar, F. Huang.
   NeurIPS Datasets and Benchmarks Track, 2024.
6. **Towards Mitigating Hallucinations in Large Vision-Language Models by Refining Textual Embeddings** Link
   A. Agrawal, G. KV, R. Aralikatti, G. Jagatap, J. Yuan, V. Kamarshi, A. Fanelli, F. Huang.
   ACL 2026; ICLR 2026 Workshop (MM Intelligence).
7. **Scheduling Thoughts: Learning the Order of Thought in Diffusion Language Models** Link
   J. Xu*, M. Liu*, A. Agrawal, Y. Chen, F. Huang.
   ICML 2026.
8. **WAVES: Benchmarking the Robustness of Image Watermarks** Link
   B. An*, M. Ding*, T. Rabbani*, A. Agrawal, Y. Xu, C. Deng, S. Zhu, A. Mohamed, Y. Wen, T. Goldstein, F. Huang.
   ICML 2024.
9. **A Technical Report on "Erasing the Invisible": The 2024 NeurIPS Competition on Stress Testing Image Watermarks** Link
   M. Ding*, B. An*, T. Rabbani*, C. Deng*, A. Satheesh, S. Chakraborty, M. Saberi, Y. Wen, K. R. Sang, A. Agrawal, X. Zhao, M. Zhou, M. A. Hartley, L. Li, Y. X. Wang, V. M. Patel, S. Feizi, T. Goldstein, F. Huang.
   NeurIPS 2025.
10. **PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models** Link
    M. Panaitescu-Liess, P. Pathmanathan, Y. Kaya, Z. Che, B. An, S. Zhu, A. Agrawal, F. Huang.
    SafeGenAi@NeurIPS 2024; NAACL 2025.

11. **Robustness to Multi-Modal Environment Uncertainty in MARL using Curriculum Learning** Link
A. Agrawal, R. Aralikatti, Y. Sun, F. Huang.
MASEC Workshop at NeurIPS 2023.

12. **Revisiting Parameter Sharing in Multi-Agent Deep Reinforcement Learning** Link
J. K. Terry, N. Grammel, S. Son, B. Black, A. Agrawal.
In review at TMLR.

13. **Rtaw: An Attention Inspired Reinforcement Learning Method for Multi-Robot Task Allocation in Warehouse Environments** Link
A. Agrawal, A. S. Bedi, D. Manocha.
ICRA 2023.

14. **DC-MRTA: Decentralized Multi-Robot Task Allocation and Navigation in Complex Environments** Link
A. Agrawal, S. Hariharan, A. S. Bedi, D. Manocha.
IROS 2022.

15. **Learning when to Trust which Teacher for Weakly Supervised ASR** Link
A. Agrawal, M. Rao, A. K. Sahu, G. Chennupati, A. Stolcke.
INTERSPEECH 2023.

EMPLOYMENT

Capital One Nov 2024 – May 2025
Ph.D. Research Intern McLean, VA

- Defined and mitigated bias in process reward model leading to reward hacking using a novel loss, and hard-negative data generation. [1]
- Application: Robustness to reward hacking for customer transaction data. Received Oral Recommendation at ICLR AFAA Workshop.

Dolby Laboratories May 2024 – Aug 2024
Ph.D. Research Intern San Francisco, CA

- Used interpretability analysis to identify and mitigate attention imbalance as a root cause of hallucinations in vision-language models. [6]

Amazon Alexa AI May 2022 – Oct 2022
Applied Scientist Intern Sunnyvale, CA

- Developed unsupervised methods for improving speech recognition models using expert feedback. [15]

Indian Institute of Science Aug 2019 – Aug 2020
Research Assistant with Prof. Debasish Ghose Bangalore, India

CAMMA Group, University of Strasbourg Jan 2019 – Jul 2019
Undergraduate Thesis Researcher with Prof. Nicolas Padoy France

SKILLS

Python, C, C++, PyTorch, Hugging Face, Transformers, Accelerate, DeepSpeed, TensorFlow, NumPy, VeRL, TRL, OpenRLHF, vLLM, Process Reward Model, reasoning LLMs, VLMs, diffusion models, multi-agent systems, robot learning.

HONORS & AWARDS

- Outstanding Achievement Award, CS Department, University of Maryland, 2025–2026.
- Travel Grant: ICRA 2023, GoldHaber, ICSSA, and Braid Scholar 2020.
- Dean's Fellowship, University of Maryland, 2022.
- Undergrad Award: Best Graduating Girl Student Award, 2019.
- Undergrad Award: Spirit of Invention 'InvEnt' Scholarship Award, Avery Dennison Foundation given to 6 students in a batch.
- Undergrad Award: 100% tuition waiver for 3 out of 4 years for GPA $\sim$10.
- JASSO Scholarship for a summer internship in Japan during undergraduate studies.

SERVICES

Reviewer: NeurIPS, ICML, ICLR, IROS, ACL ARR, AAAI, I-RAL, ICRA.
Organizer: Watermark Competition at NeurIPS 2024.
Teaching: Randomized Algorithms, Machine Learning, and Robot Programming at UMD; Neural Network and Fuzzy Logic at BITS Pilani.