Roozbeh Nahavandi

Department of Computer Science and Engineering, Ohio State University, Columbus, OH
nahavandi.2@osu.edu | Website | LinkedIn | GitHub

SUMMARY

PhD student at Ohio State researching AI safety and alignment: how LLMs make decisions under trade-offs, against other agents, and under chance, and what RL post-training changes about that. Moving toward LLM agents, with a focus on RL training and strategic decision-making. Combines RL training, behavioral evaluation, and interpretability.

EDUCATION

Ohio State University **Aug. 2023 – Present**
Ph.D. in Computer Science and Engineering, GPA: 3.85/4.0 *Columbus, OH*
- Advisor: Dr. Andrew Perrault

University of Tehran **Sep. 2018 – June 2023**
B.Sc. in Electrical Engineering, GPA: 3.77/4.0 *Tehran, Iran*
- Best Undergraduate Project Award, School of Electrical and Computer Engineering

PUBLICATIONS

AfterDeleteBench: Auditing Post-Deletion Information Persistence in LLM Agents | arXiv
- Ehsan Saleh, **Roozbeh Nahavandi**, Sama Salarian, Raef Bassily, Andrew Perrault. *arXiv preprint, 2026.*
- Behavioral audit showing that deleting a record from an LLM agent's memory does not remove the information: deleted secrets are recovered in about two thirds of cases and rebuilt from stored fragments in 62–99% across Qwen, Mistral, and Phi-3.

Many Preferences, Few Policies: Compact Portfolios for Multi-Objective LLM Alignment | arXiv
- Cheol Woo Kim*, Jai Moondra*, **Roozbeh Nahavandi**, Andrew Perrault, Milind Tambe, Swati Gupta. *Under review at ICLR 2027.*
- Designed and executed all experiments for PALM, a provably near-optimal portfolio of aligned LLMs for users with heterogeneous preferences; built the fine-tuning and evaluation pipelines across math reasoning, safety, and helpfulness.

RESEARCH EXPERIENCE

Graduate Research: LLM Alignment and RL Post-Training **Aug. 2023 – Present**
Advisor: Dr. Andrew Perrault *Ohio State University*
- Evaluating the behavioral effects of safety training on LLM agents in autonomous, multi-agent environments.
- Built and ran LLM fine-tuning and RL post-training pipelines end to end; developing constrained optimization methods for multi-objective RLHF that enforce per-objective performance thresholds instead of fixed reward weights.
- Training LLMs with self-play RL on stochastic games, studying transfer to general reasoning and credit assignment under chance.
- Studying how RL and supervised fine-tuning shift persona and trait representations in LLMs.

TECHNICAL SKILLS

- **RL & Alignment**: RL post-training (GRPO, PPO, RLVR), constrained and multi-objective RLHF, self-play RL, credit assignment, reward modeling, behavioral evaluation of LLMs, interpretability
- **Frameworks**: PyTorch, Hugging Face (Transformers, TRL, Datasets), veRL, DeepSpeed, Weights & Biases
- **Programming**: Python, C/C++, Bash, LaTeX
- **Infrastructure**: Slurm HPC, multi-GPU distributed training, large-scale experiment management

SERVICE & TEACHING

- **Reviewer**: NeurIPS 2026, ICML 2026, ICLR 2025
- **Teaching Assistant**, Ohio State CSE: Artificial Intelligence (Spring 2024), Computer Networking (Fall 2023)