Yaswanth Chittepu

PhD Student, Computer Science, University of Massachusetts Amherst · Amherst, MA

ychittepu@umass.edu | Website | GitHub | LinkedIn | Google Scholar

Research Interests

My research focuses on designing **safe and efficient reinforcement learning algorithms**: alignment of large language models (LLMs) with high-confidence guarantees and risk control, reward overoptimization, offline RL, imitation learning, inverse RL, and LLM agents.

Education

University of Massachusetts Amherst — *Amherst, MA, USA*
Doctor of Philosophy, Computer Science | GPA: 3.97/4.0 — *Expected 2027*

- Advised by **Prof. Scott Niekum** (SCALAR Lab)

University of Massachusetts Amherst — *Amherst, MA, USA*
Master of Science, Computer Science | GPA: 3.97/4.0 — *2021*

Indian Institute of Technology (IIT), Kharagpur — *Kharagpur, India*
Bachelor of Technology, Mechanical Engineering — *2016*

Publications

* Equal contribution.

- *Convex-Concave Reinforcement Learning*
  Shripad V. Deshmukh, **Yaswanth Chittepu**, Dhawal Gupta, Philip S. Thomas, Scott Niekum | Preprint 2026 [paper] [code]
- *TRIBE: Trust-Region Policy Improvement from Batch Experience*
  Yaswanth Chittepu*, Shripad V. Deshmukh*, Dhawal Gupta*, Philip S. Thomas, Scott Niekum | Under review, 2026
- *Learning Compact State Abstractions for LLM Agents*
  Yaswanth Chittepu, Raghavendra Addanki, Anup Rao, Branislav Kveton, Scott Niekum | Under review, 2026
- *Safe Inference-Time Alignment via Lagrangian Reward Augmentation*
  Yaswanth Chittepu, Ativ Joshi, Sohini Chintala, Scott Niekum | Conference on Language Modeling (COLM) 2026 [paper] [website]
- *Safe RLHF Beyond Expectation: Stochastic Dominance for Universal Spectral Risk Control*
  Yaswanth Chittepu, Ativ Joshi, Rajarshi Bhattacharjee, Scott Niekum | Preprint 2026 [paper]
- *ML-Tool-Bench: Tool-Augmented Planning for ML Tasks*
  Yaswanth Chittepu, Raghavendra Addanki, Tung Mai, Anup Rao, Branislav Kveton | Preprint 2025 [paper]
- *Adaptive Margin RLHF via Preference over Preferences*
  Yaswanth Chittepu, Prasann Singhal, Greg Durrett, Scott Niekum | Preprint 2025 [paper]
- *Reinforcement Learning from Human Feedback with High-Confidence Safety Guarantees*
  Yaswanth Chittepu*, Blossom Metevier*, Will Schwarzer, Austin Hoag, Scott Niekum, Philip S. Thomas | Reinforcement Learning Conference (RLC) 2025; Reinforcement Learning Journal, vol. 6 [paper] [website] [code]
- *Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms*
  Rafael Rafailov*, **Yaswanth Chittepu***, Ryan Park*, Harshit Sikchi*, Joey Hejna*, W. Bradley Knox, Chelsea Finn, Scott Niekum | NeurIPS 2024; also at the Models of Human Feedback for AI Alignment (MHFAIA) workshop @ ICML 2024 and the RLBrew workshop @ RLC 2024 [paper]

Research Experience

Adobe Research | Research Intern — *San Jose, CA, USA*
Compact state abstractions for LLM agents — *May 2026 – Aug 2026*

- Formulated agent state for black-box LLM agents as a **compact abstraction of the interaction history** that preserves a fixed reference agent's task performance, connecting LLM agents to classical state abstraction.
- Represented state abstractions as **deterministic recursive programs** and discovered them with an **LLM-guided genetic algorithm**; the learned programs are **directly inspectable** and need no extra model calls at deployment.
- On ALFWorld and ScienceWorld, reduced mean policy input by up to 71% and 57% while preserving near-reference task success, exceeding full-history success on 4 of 10 task families and staying within 8 points on the rest.

Adobe Research | Research Intern *San Jose, CA, USA*

TOOL-AUGMENTED PLANNING FOR ML AGENTS *May 2025 – Aug 2025*

- Built **ML-Tool-Bench**, a benchmark for tool-augmented ML agents with 61 specialized tools and 15 Kaggle tasks.
- Designed planning methods using **shaped deterministic rewards** with structured textual feedback, and **task decomposition** into sub-tasks, improving over ReAct baselines by 16.5 leaderboard percentile points.
- Introduced **in-memory object management** (a shared scratchpad of named artifacts) to improve agent planning and workflow reliability.

Robotics Research Center, IIIT Hyderabad | Research Intern *Hyderabad, India*

REINFORCEMENT LEARNING FOR HUMANOID MANIPULATION *Jul 2018 – Jul 2019*

- Learned reaching and **dual-arm manipulation policies** for a humanoid with **DDPG, modified for shared action spaces** to capture coupling between motors; reached goals among obstacles with an 85% success rate, with cubic-spline interpolation for smooth, physically executable trajectories.
- Trained a humanoid to **imitate human motion from videos**: pose estimation with Human Mesh Recovery, motion reconstruction to remove jitter and infeasible poses, and PPO in MuJoCo with a pose-similarity reward.

Research Projects

TRIBE: Trust-Region Policy Improvement from Batch Experience *SCALAR Lab, UMass Amherst*

UNDER REVIEW 2026

- Formulated LLM post-training from scalar rewards as policy improvement in **importance-ratio space**: Stage 1 solves a **finite-dimensional convex program** over the sampled responses' likelihood ratios under an f-divergence trust region, and Stage 2 fits the model to them with weighted behavior cloning; a χ^2 trust region recovers GRPO-style advantages.
- One method for **on-policy, off-policy, and fully offline** training, including data from unavailable or closed-source generators; competitive with strong on-policy baselines on GSM8K and MATH, and the only evaluated method to improve over rejection fine-tuning on UltraFeedback in both win rate and reward advantage.

Safe Inference-Time Alignment (LARA) *SCALAR Lab, UMass Amherst*

COLM 2026

- Moved the Safe RLHF constraint **from training to decoding**: dualizing the constraint reduces the search over policies to a **one-dimensional convex problem** over a **single Lagrange multiplier** $\lambda^\star$.
- Calibrated $\lambda^\star$ by binary search on a small prompt set in about two minutes; the resulting reward $r - \lambda^\star c$ plugs into Best-of-N, ARGS, PARGS, or CD-Q with **no weight updates**, and Best-of-N **approaches finetuned baselines**.

High-Confidence Safe RLHF (HC-RLHF) *SCALAR Lab, UMass Amherst*

RLC 2025

- Proposed the **first Seldonian algorithm for RLHF**: optimize helpfulness under a pessimistic upper-confidence bound on cost, then return the model only if it passes a **held-out statistical safety test**.
- Derived a **variance-aware policy gradient** for candidate selection, **proved** that an unsafe model is returned with probability at most δ, and validated on Qwen2-1.5B, Qwen2.5-3B, and Llama-3.2-3B with improved harmlessness and helpfulness.

Risk-Sensitive Alignment beyond Expected Cost *SCALAR Lab, UMass Amherst*

PREPRINT 2026

- Replaced the expected-cost constraint in Safe RLHF with **first-order stochastic dominance** over the full cost distribution, which controls a broad family of **spectral risk measures**; made it tractable with optimal transport and Sinkhorn iterations.

Adaptive Margin RLHF via Preference over Preferences *SCALAR Lab, UMass Amherst*

PREPRINT 2025

- Proposed **DPO-PoP**, which uses preference-over-preference supervision to infer **adaptive margins**, **outperforming fixed and ground-truth margin baselines** on UltraFeedback; analyzed the trade-off between discriminative and generative performance and designed sampling strategies to prioritize either.

Reward Overoptimization in Direct Alignment Algorithms *SCALAR Lab, UMass Amherst*

NEURIPS 2024

- Showed that direct alignment algorithms (DPO, IPO, SLiC) **degrade at high KL budgets** like classical RLHF, **begin overfitting before a single epoch completes**, and, being under-constrained, often **favor out-of-distribution responses**; smaller models overfit to simple features such as length.

Model-Free RL for Diabetes Treatment *UMass Amherst*

M.S. PROJECT

- Learned **insulin-dosing policies** from continuous glucose monitoring data with DDPG, using **high-confidence off-policy improvement** to switch from the behavior policy only when the new policy is better with high confidence.

Industry Experience

athenahealth | Member of Technical Staff *Remote*

PAYMENTS INFRASTRUCTURE *Jul 2021 – Mar 2022*

- Designed and developed the **payments infrastructure** for bill generation and customer payment processing; led the database schemas and APIs for backend–frontend integration.

athenahealth | Software Engineering Intern *Remote*

MENTOR MATCHING WEB APP *Jun 2020 – Aug 2020*

- Built a React web app that matches mentors to mentees under preferences and constraints with an **optimized backtracking algorithm**, REST APIs, and SQL schemas; deployed it on **AWS** via a **Jenkins CI/CD pipeline** (Docker).

dunnhumby | Data Scientist *Gurugram, India*

COMPUTER VISION AND CUSTOMER ANALYTICS *Jun 2016 – Jun 2018*

- Automated text extraction from scanned receipts: **conditional GANs** to detect tables, and Hough transforms to extract bounding boxes and key–value pairs.
- Segmented Tesco stores in Central Europe with **Gaussian mixture models** (EM, PCA, BIC for model selection), improving silhouette score by over 30% versus k-means.

Talks and Presentations

- *Safe Inference-Time Alignment via Lagrangian Reward Augmentation* | Poster, COLM 2026, San Francisco | Oct 2026
- *Reinforcement Learning from Human Feedback with High-Confidence Safety Guarantees* | Talk and poster, RLC 2025, Edmonton | Aug 2025
- *Scaling Laws for Reward Model Overoptimization in Direct Alignment Algorithms* | Poster, NeurIPS 2024, Vancouver | Dec 2024

Awards

2026 Gold Reviewer, International Conference on Machine Learning (ICML)

Service

- **Conference reviewer:** NeurIPS 2026, ICML 2026 (Gold Reviewer), ICLR 2026, COLM 2026.
- **Holistic reviewer:** Reinforcement Learning Conference (RLC) 2026.
- **Journal reviewer:** Transactions on Machine Learning Research (TMLR).
- **Workshop reviewer:** RLBrew @ RLC 2024, 2025; Models of Human Feedback for AI Alignment @ ICML 2024, 2025.

Skills

Languages	Python, C, C++, JavaScript/TypeScript, MATLAB
Methods	RLHF (PPO, GRPO, RLOO, DPO), supervised fine-tuning (SFT), distributed multi-GPU training, LLM agents and tool use
ML / RL	PyTorch, JAX, Hugging Face Transformers, TRL, DeepSpeed, Accelerate, vLLM, LangGraph, Gymnasium, Stable-Baselines
Robotics	ROS, MuJoCo, Gazebo