

Souradip Chakraborty

Postdoctoral Researcher at MIT CSAIL

Cambridge, Massachusetts, USA
301-204-3898 ■ ✉ soura24@mit.edu

Summary

Currently a **Postdoctoral Researcher at MIT CSAIL**, working with **Prof. Omar Khattab** on the foundations of reinforcement learning for AI agents. My research focuses on **Pedagogical Reinforcement Learning**, **Foresight Reinforcement Learning**, post-training of language models, AI alignment, and long-horizon agentic systems, combining theoretical foundations with practical algorithms for continually learning AI agents.

I completed my Ph.D. in Computer Science at the **University of Maryland, College Park**, under **Prof. Furong Huang** and **Prof. Dinesh Manocha**, while collaborating closely with **Prof. Mengdi Wang** (Princeton University), **Prof. Amrit Singh Bedi** (University of Central Florida), and **Prof. Aldo Pacchiano** (Boston University) on foundations of reinforcement learning for Post-training and alignment

I am a recipient of the **University of Maryland Computer Science Certificate of Outstanding Achievement**, **Outstanding Paper Awards** at NeurIPS workshops, and **Outstanding Reviewer Awards** at NeurIPS and AISTATS, with my research receiving over 1,600 citations.

Work Experience

Massachusetts Institute of Technology

Postdoc Researcher, Advisor: Prof. Omar Khattab

July 2026 – Present

- Foundations of Pedagogical RL for Post-training LLM agents

Google Research

Student Researcher, Mentor: Dr. Rahul Kidambi, Alekh Agarwal, Chris Dann

Aug 2025 – Dec 2025

- Fundamental research to improve Credit Assignment in Large Reasoning Models

Google Research

Student Researcher, Mentor : Dr. Ahmad Beirami

Oct 2024 – May 2025

- Fundamental research to improve Reasoning and Alignment of AI Agents

JP Morgan Chase AI Research

Research Intern, Mentor : Dr. Sumitra Ganesh

May 2024 – Aug 2024

- Controlled Decoding using Mixture of Agents Alignment (ICLR'2025)

Walmart Labs

Data Scientist, Mentor : Dr. Saswata Sahoo

Feb 2018 – Aug 2021

- Fundamental research improving Multi-modal Recommendation and Retrieval Systems (Patent)

Education

Massachusetts Institute of Technology

Postdoc, Computer science

Fall 2026 – Present

Advisor: Prof. Omar Khattab

Research: Foresight RL for Post-training LLM agents

University of Maryland, College Park

PhD, Computer science (Dean's Fellowship)

Fall 2021 – Fall 2026

Advisor: Prof. Furong Huang, Prof. Dinesh Manocha

Mentor: Prof. Mengdi Wang (Princeton University), Dr. Mohammad Ghavamzadeh (Amazon AGI)

Research: A Principled Approach to AI Alignment for Agents: Theoretical Foundations and Practical Algorithms

Indian Statistical Institute, Bangalore

Master's Degree (summa cum laude)

2016 – 2018

MS with Major in Data Science & Machine Learning **GPA: 9.4/10; Rank — 1 (Batch Topper)**

Advisor: Dr. B. S. Daya Sagar

Courses: Probability, Statistics, Inference, Statistical Machine Learning, Design of Experiments, Optimization and Reliability Theory.

Preprints

- Souradip Chakraborty, Soumya Suvra Ghosal, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, Amrit Singh Bedi **“Does Thinking More always Help? Understanding Test-Time Scaling in Reasoning Models”** Ref :arxiv
- Bhrij Patel, Souradip Chakraborty, Wesley A Suttle, Mengdi Wang, Amrit Singh Bedi, Dinesh Manocha **“AIME: AI System Optimization via Multiple LLM Evaluators”** Ref :arxiv
- Souradip Chakraborty, Mucong Ding, Vibhu Agrawal, Zora Che, Alec Koppel, Mengdi Wang, Amrit Bedi, Furong Huang **“SAIL: Self-Improving Efficient Online Alignment of Large Language Models”** Ref :arxiv
- Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, Sayak Ray Chowdhury **“Active Preference Optimization for Sample Efficient RLHF”** Ref :arxiv
- Xiyang Wu, Souradip Chakraborty, Ruiqi Xian, Jing Liang, Tianrui Guan, Fuxiao Liu, Brian M. Sadler, Dinesh Manocha, Amrit Singh Bedi **“On the Safety Concerns of Deploying LLMs/VLMs in Robotics: Highlighting the Risks and Vulnerabilities”** Ref :arxiv

Publications

- Souradip Chakraborty, Sujay Bhatt, Udari Madhushani Sehwag, Soumya Suvra Ghosal, Jiahao Qiu, Mengdi Wang, Dinesh Manocha, Furong Huang, Alec Koppel, Sumitra Ganesh **“Collab: Controlled Decoding using Mixture of Agents for LLM Alignment”** *The Thirteenth International Conference on Learning Representations (ICLR 2025)*
- Soumya Suvra Ghosal*, Souradip Chakraborty*, Vaibhav Singh, Tianrui Guan, Mengdi Wang, Ahmad Beirami, Furong Huang, Alvaro Velasquez, Dinesh Manocha, Amrit Singh Bedi **“Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment”** *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 (CVPR 2025)*
- Pankayaraj Pathmanathan, Souradip Chakraborty, Xiangyu Liu, Yongyuan Liang, Furong Huang **“Is poisoning a real threat to LLM alignment? Maybe more so than you think”** *The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI 2025)*
- Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, Furong Huang **“Transfer Q Star: Principled Decoding for LLM Alignment”** *Thirty-Eighth Annual Conference on Neural Information Processing Systems (Neurips 2024)*
- Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Furong Huang, Dinesh Manocha, Amrit Singh Bedi, Mengdi Wang **“MaxMin-RLHF: Towards Equitable Alignment of Large Language Models with Diverse Human Preferences”** *41st International Conference on Machine Learning (ICML 2024)*
- Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, Furong Huang **“On the Possibilities of AI-Generated Text Detection”** *41st International Conference on Machine Learning (ICML 2024)*
- Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Dinesh Manocha, Furong Huang, Mengdi Wang **“Principal-Driven Reward Design and Agent Policy Alignment via Bilevel-RL”**, *Twelfth International Conference on Learning Representations (ICLR 2024)*
- Xiangyu Liu, Souradip Chakraborty, Yanchao Sun, Furong Huang **“Rethinking Adversarial Policies: A Generalized Attack Formulation and Provable Defense in Multi-Agent RL”** *Twelfth International Conference on Learning Representations (ICLR 2024)*
- Souradip Chakraborty*, Soumya Suvra Ghosal*, Jonas Geiping, Furong Huang, Dinesh Manocha, Amrit Bedi **“A Survey on the Possibilities & Impossibilities of AI-generated Text Detection”** *Transactions on Machine Learning Research (TMLR 2024)*

- Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Mengdi Wang, Furong Huang, Dinesh Manocha **“STEERING: Stein Information Directed Exploration for Model-Based Reinforcement Learning”**, *40th International Conference on Machine Learning (ICML 2023)*
- Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Dinesh Manocha, Furong Huang, Mengdi Wang **“Principal-Driven Reward Design and Agent Policy Alignment via Bilevel-RL”**, *Interactive Learning with Implicit Human Feedback Workshop (ICML 2023)*
- Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Brian M. Sadler, Furong Huang, Pratap Tokekar, Dinesh Manocha **“Posterior Coreset Construction with Kernelized Stein Discrepancy for Model-Based Reinforcement Learning”**, *Thirty-Seventh AAAI Conference on Artificial Intelligence, (AAAI 2023, Oral)*[pdf]
- Souradip Chakraborty, Amrit Singh Bedi, Amrit Singh Bedi, Alec Koppel, Pratap Tokekar, Dinesh Manocha **“Dealing with Sparse Rewards in Continuous Control Robotics via Heavy-Tailed Policy Optimization”**, *2023 International Conference on Robotics and Automation, (ICRA 2023)*[pdf]
- Souradip Chakraborty*, Kasun Weerakoon*, Nare Karapetyan, Adarsh Jagan Sathyamoorthy, Amrit Singh Bedi, Dinesh Manocha **“HTRON:Efficient Outdoor Navigation with Sparse Rewards via Heavy Tailed Adaptive Reinforce Algorithm”**, *Conference on Robot Learning CoRL 2022*[pdf]
- Xiangyu Liu, Souradip Chakraborty, Furong Huang **“Controllable Attack and Improved Adversarial Training in Multi-Agent Reinforcement Learning”**, *Outstanding Paper Award at TSRML, Neurips’2022*[pdf]
- Amrit Singh Bedi, Souradip Chakraborty, Anjaly Parayil, Brian Sadler, Pratap Tokekar, Alec Koppel **“On the Hidden Biases of Policy Mirror Ascent in Continuous Action Spaces”**, *39th International Conference on Machine Learning (ICML 2022, Spotlight)*
- Souradip Chakraborty*, Ekansh Verma*, Ryan **“High-Dimensional Bayesian Optimization with Invariance”**, *ICML 2022 Workshop on Adaptive Experimental Design and Active Learning in the Real World* [pdf]
- Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott and Francesco Mosconi **“BioMed-BERT: A Pre-trained Biomedical Language Model for QA and IR”**, *28th International Conference on Computational Linguistics (COLING’2020)* [Website: pdf]
- Saswata Sahoo, Souradip Chakraborty **“Graph Spectral Feature Learning for Mixed Data of Categorical and Numerical Type”**, *25th International Conference on Pattern Recognition (ICPR’2020)* [pdf]
- Souradip Chakraborty, Sayak Paul, Aritra Roy Gosthipaty **“G-SimCLR: Self-Supervised Contrastive Learning with Guided Projection via Pseudo Labelling”**, *IEEE ICDM’2020, DLKT Workshop Proceedings* [pdf]
- Souradip Chakraborty, Ekansh Verma, Saswata Sahoo, Jyotishka Datta **“FairMixRep: Self-supervised Robust Representation Learning for Heterogeneous Data with Fairness constraints”**, *IEEE ICDM’2020, DLC Workshop Proceedings* [pdf]
- Ekansh Verma, Souradip Chakraborty, Vinodh Kumar **“Propaganda Fragment Detection using Diversified BERT Architectures based Ensemble Learning”**, *SemEval Workshop Proceedings, COLING’2020* [pdf]
- Ekansh Verma, Souradip Chakraborty, Vinodh Kumar **“Deep Multi-level Fusion Learning Framework for Multi-modal Product Classification”**, *SIGIR ecom’2020 Workshop Proceedings* [pdf]

- Ojaswini Chhabra, Souradip Chakraborty “**Siamese Triple Ranking Convolution Network in Signature Forgery Detection**”, selected at AICAAM’19, selected and submitted to Pertanika journal [pdf]
- Soumya Dasgupta, Kaushik Halder, Shohan Banerjee Souradip Chakraborty, Amitava Gupta “**Stability analysis and controller synthesis of networked control system (NCS) with arbitrary packet drop-outs**” 2nd International Conference on Electronics and Communication Systems (ICECS)’2015 (Link : paper)

Books & Live Projects

- Souradip Chakraborty, Sayak Paul “**Extractive Text Summarization of News Articles with NLP, Deep Learning, Python - An attention-based framework**”, *Manning Books Live Project* (Link: Website)

Patents

- Gregory Dixon, Souradip Chakraborty, Ojaswini Chhabra, Mallikharjuna Mv “**Reverse Engineering Food Ingredient Share estimation using Constrained Optimization**”, US Patent, Walmart Ref. 6031US01.
- Souradip Chakraborty, Abhishek Mishra, Somedip Karmakar “**Systems and methods for Unsupervised image processing**”, US Patent, Ref. US11688049B2
- Pranay Dugar Souradip Chakraborty “**Automated planogram anomaly detection with Computer Vision**”, US Patent, Ref. US11669843B2
- Souradip Chakraborty, Mani Garlapati “**Retail Based Cost Reverse Engineering and Cost comparison within Item Similarity Clusters for Cost Negotiations**”, US Patent, Walmart Ref. 5928US01.
- Souradip Chakraborty, Mani Garlapati “**Retail Based Cost Reverse Engineering and Cost comparison within Item Similarity Clusters for Cost Negotiations**”, US Patent, Walmart Ref. 5928US01.
- Souradip Chakraborty, Mani Garlapati “**Systems and Methods for Identifying Negotiable Items (Cost Analytics)**”, US Patent, Walmart Ref. 5604US01.
- Souradip Chakraborty, Rajesh Shreedhar Bhat, Mani Garlapati, “**System and Method For Automated Electronic Catalogue Management and Image Quality Assessment**”, US Patent, Walmart Ref. 5118US01.
- Souradip Chakraborty, Rajesh Bhat, Mani Garlapati, Lakshmi Praneetha Kommuru, “**Generating Customized Alerts with Computer Vision and Machine Learning**”, US Patent, Walmart Ref. 5008US01.
- Souradip Chakraborty, Mani Garlapati “**Architecturally-Distributed Apparatus and Method to Form and Leverage Clustered Content (Customer intent based recommendation system)**”, US Patent, Walmart Ref. 4970US01.
- Souradip Chakraborty, Ojaswini Chhabra “**System and Method for Detecting Signature Forgeries**”, US Patent, Walmart Ref. 5603US01.
- Souradip Chakraborty, Rajesh Shreedhar Bhat, Somedip Karmakar “**Systems and methods for inventory management with Deep Vision**”, US Patent, Ref. US20230020026A1.
- Somedip Karmakar, Souradip Chakraborty, Abhishek Mishra “**Image processing based methods and apparatus for planogram compliance**”, US Patent, Ref. US20220415012A1.

Selected Honors & Awards

- **Outstanding Paper Award** at Neurips’2024, The 3rd New Frontiers in Adversarial Machine Learning
- **Outstanding Paper Award** at Neurips’2022, TSRML Workshop for our research on Adversarial Robustness in Multiagent Reinforcement Learning
- **Outstanding Reviewer Award** Neurips’2023 || Awarded to 5% of the reviewers’
- **Outstanding Reviewer Award** Neurips’2022 || Awarded to 5% of the reviewers’
- **Outstanding Reviewer Award** AISTATS’2023 || Awarded to 2% of the reviewers’
- **Dean’s Fellowship** Awarded to only 5 students in the first and second year in the Computer Science department at the University of Maryland.
- **Conference Awards** Received Travel and Participant awards for ICML’23, AISTATS’23, Neurips’22.
- 2019: Selected as the Technical speaker for the very prestigious **Data Hack Summit’2019 on Captioning & Attention Models by Analytics Vidhya**.
- 2019: Key-Note Lecturer to the Faculty of Presidency University, Bangalore in Statistical Learning Theory **Faculty Development Program**.

- 2019: Key-Note Lecturer to the students of Computer Science department of Coimbatore Institute of Technology **Machine Learning Workshop with Python**.
- 2019: Invited as a Technical Keynote speaker for **Target Talks AI Session-3** Bangalore'19.
- 2018: Selected as the Youngest Technical Keynote speaker for the very prestigious **Data Hack Summit'2018 by Analytics Vidhya**.
- 2018: **Batch Topper Certification** and **Endowment** for the highest academic performance (Rank-1) in Master's, Indian Statistical Institute.
- 2017: Selected at **Novartis Biocamp 2017** and represented ISI Bangalore in Novartis as a Data Scientist (top 50 nationwide).
- 2010-2014: **4-year scholarship** for academic excellence, Ministry of Human Resource & Development, India

Blogs & Research Articles

- Souradip Chakraborty "**Detection of COVID-19 presence from Chest X-ray scans using CNN & Class Activation Maps**", *Towards Data Science, Medium'2020*
- Souradip Chakraborty, Rajesh Shreedhar Bhat "**Reducing the Carbon Foot Prints of CNNs at the cost of interactions-Depthwise & Pointwise Convolution**", *Towards Data Science, Medium'2020*
- Souradip Chakraborty "**Bayesian Thinking for Linear Regression @ Kaggle Days Meetup**", *Towards Data Science, Medium'2020*
- Souradip Chakraborty, Amlan Jyoti Das & Sai Yashwanth "**Risks and Caution on applying PCA for Supervised Learning Problems**", *Towards Data Science, Medium'2019*
- Souradip Chakraborty, Rajesh Shreedhar Bhat "**Why not Mean Squared Error(MSE) as a loss function for Logistic Regression?**", *Towards Data Science, Medium'2019*.
- Souradip Chakraborty "**Dimensionality Reduction in Supervised Framework and Partial Least Square Regression**", *Analytics Vidhya, Medium'2019*.

Competitions

- **Runners up, Codeception 2019** Walmart Labs International Hackathon.
- **Rank-13, Crowd Analytix's Propensity to Fund Mortgages competition 2019** – Implemented LightGBM with error analysis to identify the curvature of the variables and interaction among the features in modelling the response variable.
- **Bronze Medal, Capillary Machine Learning Hackathon by Analytics Vidhya'2019** – Implemented Alternating Least Squares Method for Implicit recommendation.
- **Bronze Medal, WNS Analytics Wizard 2018 challenge** – An ensemble of Boosting and Deep Neural nets with synthetic minority oversampling was implemented to solve the classification problem with class imbalance.

Skillsets

- **Programming Languages:** Python, R, C/C++, MATLAB; **Web Development:** HTML
- **Deep Learning components:** *RNNs, *CNNs, *GANs, Attention, Capsule (* denotes variants)
- **Frameworks/Databases:** PyTorch, Keras, Tensorflow, PySpark, Teradata, MongoDB, Hive, SQL
- **Tools/Softwares:** NLTK, OpenCV, Octave, Docker, $\text{\LaTeX}$