

Pragya Srivastava

Pre-doctoral Researcher, Google DeepMind

@pragya8srivastava@gmail.com Github Google Scholar +91-7617478783

Education

Indian Institute of Technology, Delhi (IIT D)

July 2019 - May 2023

B.Tech in Engineering Physics; GPA: 8.7/10.0

Research Experience

Google DeepMind

July 2024 - Present

Pre-doctoral Researcher | Advisors: Dr. Aravindan Raghuv eer, Dr. Karthikeyan Shanmugam, Prof. Doina Precup

Projects: Causally Robust Reward Models [S.1], Multi-objective LLM Alignment, Sample Efficient RL for Post-training Gemini Models

University of Texas, Austin

March 2024 - June 2024

Research Intern | Advisors: Peihao Wang, Prof. Atlas Wang

Project: Theoretical Analysis of the bottlenecks of Structured State Space Models [C.3]

Microsoft Research, India

May 2022 - February 2024

Research Intern | Advisors: Dr. Amit Sharma, Dr. Amit Deshpande, Tanuja Ganu

Project: Multi-hop reasoning over semi-structured data [C.1], In-context Learning [C.2], Active Learning for Reward Models [C.4]

ShareChat AI, India


Feb 2023 - May 2023

Machine Learning Engineer Intern | Advisor: Srijan Saket

Project: Multimodal Feature Encoders for Enhanced Early-Stage Video Recommendations

Selected Publications


S=In Submission, C=Conference, *=Equal Contribution

[S.2] **Benchmarking Anomaly Detection for Large Language Model Alignment** 

Pragya Srivastava*, Dylan Feng*, Anca Dragan, Cassidy Laidlaw

Under Review


[Under Review at ICLR 2026]

[S.1] **Robust Reward Modeling via Causal Rubrics** 

Pragya Srivastava*, Harman Singh*, Rahul Madhavan*, Gandharv Patil, Sravanti Addepalli, Arun Suggala, Rengarajan Aravamudhan, Soumya Sharma, Anirban Laha, Aravindan Raghuv eer, Karthikeyan Shanmugam, Doina Precup

Under Review



[Under Review at ICLR 2026]

[C.4] **Outlier-Aware Preference Optimization for Large Language Models** 

Pragya Srivastava, Sai Soumya Nalli, Amit Deshpande, Amit Sharma

Bidirectional Human-AI Alignment

[ICLR'25 Bi-Align Workshop]

[C.3] **Understanding and Mitigating Bottlenecks of State Space Models through the Lens of Recency and Over-smoothing**  

Peihao Wang, Ruisi Cai, Yuehao Wang, Jiajun Zhu, Pragya Srivastava, Zhangyang Wang, Pan Li

International Conference on Learning Representations

 First-author workshop version at ICML 2024 NGSM Workshop


[ICLR'25]

[C.2] **NICE: To Optimize In-Context Examples or Not?**  

Pragya Srivastava, Satvik Golechha, Amit Deshpande, Amit Sharma

Annual Conference of the Association for Computational Linguistics

[ACL'24]

[C.1] **Evaluating LLMs' Mathematical Reasoning in Financial Document Question Answering** 

Pragya Srivastava, Manuj Malik, Vivek Gupta, Tanuja Ganu, Dan Roth

Annual Conference of the Association for Computational Linguistics

[ACL'24 Findings]

Selected Projects

Robust Reward Modeling for Mitigating Reward Hacking

February 2025 – July 2025

Advisors: Dr. Karthikeyan Shanmugam, Prof. Doina Precup

Google DeepMind

- Proposed a novel data-augmentation technique grounded in an explicit causal model to mitigate reward hacking, directly addressing failure modes in standard RLHF pipelines.
- Demonstrated **SoTA performance** on RewardBench, outperforming existing baselines while achieving significant gains in downstream alignment tasks including Iterative On-Policy DPO and Best-of-N sampling.
- Built a modular framework that standardizes data processing and training loops across diverse model families.

Multi-Objective LLM Alignment

December 2024 – February 2025

Advisors: Dr. Karthikeyan Shanmugam

Google DeepMind

- Proposed a theoretically grounded objective utilizing the Nash Social Welfare function to solve the problem of alignment under conflicting objectives, ensuring Pareto-optimal trade-offs.
- Led the end-to-end implementation of the proposed framework for Gemini Post-training, achieving superior convergence stability and **6%** metric improvements in complex domains like Math, Factuality, Safety and Tool Use.

Principled Curriculum Design for Post-training

July 2025 – Present

Advisors: Dr. Karthikeyan Shanmugam, Prof. Doina Precup

Google DeepMind

- Proposed a novel difficulty curriculum for sampling prompts with high learning signal during RL training to accelerate the training progress of a KL constrained policy.
- Achieved **3x training efficiency (FLOPs)** on smaller sized models and **ranked #1 on internal downstream benchmarks, showing significant promise for scaling to larger model sizes as well.**

Natural Language Feedback for Efficient Policy Exploration

July 2025 – Present

Advisor: Dr. Aravindan Raghuvier

Google DeepMind

- Proposed a novel technique for incorporating natural language feedback to enable efficient exploration of the solution space.
- Tested various intervention methods to optimize the trade-off between guided training speed and final model autonomy at test time.

Simulating the dynamics of Hamiltonian Systems

Aug 2022 – Nov 2022

Advisor: Prof. Vishal K. Vaibhav

B.Tech Project, IIT Delhi

- Developed a novel algorithm to infer the Energy of a Hamiltonian System and predict the trajectory of a physical system.
- Proposed a novel architecture **Recurrent Hamiltonian Neural Networks** for predicting the trajectory of physical systems with high fidelity by leveraging the **volume-preserving** property of symplectic transformations.

Few-Shot Transfer Learning for Remote-Sensing Images

Dec 2021 – April 2022

Advisor: Prof. Stefano Ermon

Stanford University

- Worked on few-shot domain adapting a model trained on **fMoW** images to **fMoW-Sentinel** images.
- Experimented with feature transformation from target to source domain using **Pix2Pix** and **Style-Transfer** methods.

Learning to decompose features into domain-specific and class-specific features

Aug 2021 – March 2022

Advisor: Prof. Pengtao Xie

UC, San Diego

- Leveraged Differentiable Neural Architecture Search (**DARTS**) for learning to decompose the representation of an input image into its domain-specific and class-specific representations.
- Proposed a **multi-level optimization** problem for adapting a model to a target domain during inference time.

Cross-Lingual Few Shot Transfer for Natural Language Inference

NLP Course Project

Advisor: Prof. Mausam

IIT Delhi

- Fine-tuned multi-lingual pre-trained models (mBERT, XLM-R) for the downstream task of inferring if one statement is an entailment, contradiction or neutral with respect to the other statement.

Named Entity Recognition in Chemical Procedures

NLP Course Project

Advisor: Prof. Mausam

IIT Delhi

- Built neural networks for sequence labelling for the purpose of extracting named entities such as reagent, action, amount, temperature, etc.
- Benchmarked **BiLSTM-CRF with multi-head self-attention** and pre-trained language models (BERT) on the provided dataset.

Selected Accolades

All India Rank in top 0.2% in JEE-Main amongst 1.5 Million candidates.
All India Rank in top 2% in JEE-Advanced amongst 150,000 candidates.
One of 12, of 20k applicants to be selected for the Google Pre-doctoral program 2024-26.

Skills

Languages	C++, Python, Java, Bash, MATLAB
Libraries	PyTorch, TensorFlow, Keras, HuggingFace, NLTK, OpenCV, Rasa, Spacy, Langchain
Tools	Kubernetes, Docker, SQL (BigQuery), WandB, Optuna, RayTune, Git, HPC (Slurm), LaTeX

References

- Prof. Doina Precup Associate Professor, McGill University
- Dr. Karthikeyan Shanmugam Senior Staff Research Scientist, Google DeepMind
- Prof. Atlas Wang Associate Professor, University of Texas, Austin
- Dr. Amit Sharma Principal Researcher, Microsoft Research India
- Dr. Amit Deshpande Principal Researcher, Microsoft Research India
- Prof. Vishal K. Vaibhav Assistant Professor, IIT Delhi
- Prof. Stefano Ermon Associate Professor, Stanford University