jatinprakash1511@gmail.com
bicycleman15.github.io
jatinprakash1511
bicycleman15
bicycleman15
NeBK8VMAAAAJ

Jatin Prakash

Research Interests

I am broadly interested in designing **scalable**, **practical** and **efficient** foundation model architectures and training algorithms: both pre-training and post-training. I believe that to accomplish this, one must tackle all the parts of the modeling stack. To this end, I have worked on: **data** [1,2], **model architectures** [3], **training algorithms** [4] and **system optimizations** [5].

Education

2024–Present **New York University**, *Courant Institute of Mathematical Sciences*
PhD in Computer Science
Advisor: Rajesh Ranganath

2018–2022 **Indian Institute of Technology Delhi (IIT Delhi)**
B.Tech (Bachelors) in Computer Science and Engineering (CSE)

Work Experience

2022-2024 **Microsoft Research**
Pre-Doctoral Research Fellow
Advisors: Manik Varma, Amit Sharma, Ramchandran Ramjee
I broadly worked on **large-scale machine learning**, specifically web-scale retrieval. I worked on two projects focusing on:
- creating scalable and fast synthetic data generation algorithms/pipelines for web-scale retrieval tasks to tackle *bad quality* click data. This algorithm scales to industrial query-ads datasets containing upto 10M+ data points [2].
- designing system optimizations to make training of large-scale retrieval models feasible and practical, bringing down the training time from weeks to just under a day [5].

Parts of above research found its way into **Microsoft Bing**, resulting in increased click yield and revenue.

Selected Research

2025 **Attention and Compression is all you need for Controllably Efficient Language Models**
Jatin Prakash, Aahlad Puli, Rajesh Ranganath
Under Review [arXiv] [Code]
ICML 2025 Efficient Systems for Foundation Models III (ES-FoMo) workshop
TLDR; We propose an architecture that provides a knob to control quality-efficiency trade-offs directly at test-time, without requiring any retraining. The proposed adaptive model outperforms efficient baselines across varying compute-memory budgets, all using a single model only. At the same time, matching dense transformer in language modeling while being upto $\mathbf{1.5-3\times}$ **faster and** $\mathbf{2-9\times}$ **memory efficient.**

2025 **What Can You Do When You Have Zero Rewards During RL?**
Jatin Prakash*, Anirudh Buvanesh*
[Notion Blog] [Code]
TLDR; We benchmarked recent RL algorithms on a simple star-graph task where they fail in zero reward scenarios, even those specially designed for this case. Turns out, a very simple data-centric intervention of just adding easy samples of the task helps unlock RL training. *Open-sourced implementations for many RL baselines (that had no official code) for the community to build upon.*

2025 **KL-Regularized Reinforcement Learning is Designed to Mode Collapse**
Anthony GX-Chen, **Jatin Prakash**, Jeff Guo, Rob Fergus, Rajesh Ranganath
Under Review [arXiv]
NeurIPS 2025 Foundations of Reasoning in Language Models (FoRLM) workshop
TLDR; We understand diversity collapse problem in RL, and how to principally fix it in 2 lines of code. Key idea is viewing KL-regularized RL as distribution matching to a target distribution. Our work explores how to define a good target for the proposal distribution (or policy in case of RL) that avoids mode collapse.

2025 **On the Necessity of World Knowledge for Mitigating Missing Labels in Extreme Classification**
Jatin Prakash*, Anirudh Buvanesh*, Bishal Santra, Deepak Saini, Sachin Yadav, Jian Jiao, Yashoteja Prabhu, Amit Sharma, Manik Varma
KDD 2025 [arXiv] [Code]
TLDR; A simple, scalable and data-centric algorithm to mitigate bad quality click-data problem in retrieval (extreme classification). This outperforms SOTA significantly, highlighting the importance of good quality dataset (that contains diverse world knowledge) for retrieval. Part of this work has been deployed in Microsoft Bing.

2023 **Renee: end-to-end training of extreme classification models**
Vidit Jain, **Jatin Prakash**, Deepak Saini, Jian Jiao, Ramachandran Ramjee, Manik Varma
MLSys 2023 [Paper] [Code]
TLDR; We unlock end-to-end training of large-scale retrieval (extreme classification) models that scales to 100M+ documents and 1B+ training examples, reducing training time from weeks to under a day. Turns out, simple end-to-end learning outperforms complicated, modular SOTA methods. This work has been deployed in Microsoft Bing.

2022 **A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Improved Neural Network Calibration**
Ramya Hebbalaguppe*, **Jatin Prakash***, Neelabh Madan*, Chetan Arora
CVPR 2022 **Oral** (top 4.2% papers) [arXiv] [Code]

Software Engineering Experience

2022 **Ivy (YC'22)**, *Graph Compiler group*, London, UK
ML Research Engineer Intern
Worked on the graph compiler that transpiles code in one ML framework to another. [Website] [GitHub]

2021 **Microsoft**, *Security and Compliance team*, IDC, Hyderabad, India
Software Engineering Intern
Worked on improving the Document Fingerprinting algorithm in M365 services for sensitive document classification.

2020 **OpenMined**, Open Source Contributions
Repository maintainer for SyferText, a privacy preserving NLP library. [Website] [GitHub]

Teaching

2022 **Introduction to Computer Science**, *Teaching Assistant*, IIT Delhi

Scholastic Achievements

- Oral presentation (top 4.2% papers) at CVPR 2022 for undergraduate thesis at IIT Delhi.
- Qualified for the ACM-ICPC Regionals 2021 programming competition.
- Secured **99th** percentile in JEE Advanced and JEE Mains 2018 examinations among a million contesting candidates.

Extra Curricular

- Competitive programming: Expert (1854) on [Codeforces], 4 stars on [Codechef].
- Core team member of the software development club of IITD, DevClub. [Github]