Mansheej Paul

✉ mansheejp@gmail.com • 🌍 mansheej.github.io

Research Interests

My research spans pre-training and post-training LLMs with a focus on optimizing data quality, distribution and curricula. I currently build synthetic data pipelines that scale inference compute to create diverse generations and develop strategies to verify and filter them into high-quality training data. Through rigorous evaluation of model behavior and how it is shaped by training data properties, I aim to create reliable, consistent and safe AI systems.

Research and Work Experience

Databricks Mosaic Research **New York City, NY**
Research Scientist *2023 – Present*
Post-training team

○ Conducted large-scale systematic experiments to analyze how data distribution, diversity and scale impact reward modeling and RL policy learning performance. Developed systems to identify and mitigate model failure modes through targeted synthetic preference data generation.
○ Developed Critique-out-Loud reward models, a method for training reward models to generate natural language critiques before scoring responses. This enables reward models to leverage variable inference compute to achieve up to a 5.8% point improvement in preference modeling accuracy on the RewardBench benchmark.

Data team

○ Led development of the **code instruction finetuning data** for *DBRX Instruct*, Databricks' foundation model, improving performance on the HumanEval code benchmark from 50% (base model) to 70% (instruct model).
○ Built the **pre-training data mixture and curriculum** for *DBRX*, Databricks' large-scale foundation model. Developed domain upsampling, a data curriculum method, to efficiently measure data mixture quality resulting in an optimized recipe that achieved an estimated **2x token efficiency** compared to MPT-7b.

FAIR CoreML, Meta AI **Menlo Park, CA**
Research Intern *2022*
Discovered the mechanistic relationship between loss landscape structure and the Lottery Ticket Hypothesis, explaining why network pruning algorithms can find high-performing sparse networks. Demonstrated that the loss basin curvature predicts the maximum achievable network sparsity that still maintains performance.

Neural Dynamics and Computation Lab **Stanford University, CA**
Ph.D. Advisor: Surya Ganguli *2018 – 2023*
Conducted research on the science of deep learning through the lens of data, loss landscapes and neural tangent kernels. Published papers at top-tier machine learning conferences on data pruning, network sparsity, and how pre-training data diversity impacts in-context learning performance.

Regulation, Evaluation, and Governance Lab **Stanford University, CA**
Research Fellow *2020 – 2021*
This work was done with partners at the Internal Revenue Service (IRS) and the Department of Labor (DOL).

○ Built an efficient data pipeline to process millions of tax returns and construct a graph representing tax relationships between businesses. Extracted features from this graph, enabling the IRS to analyze tax planning in networks of business partnerships.
○ Trained a BERT model to tag the relevant spans of text in long medical claims documents that require human review. This significantly reduced the amount of text auditors needed to evaluate for disability compensation.

Education

Stanford University **Stanford, CA**
Ph.D. in Applied Physics *2017 – 2023*
Advisor: Surya Ganguli

Brown University **Providence, RI**
B.S. in Applied Mathematics, Honors, Magna Cum Laude, Robin Truell Prize *2013 – 2017*
Honors Thesis: Random Matrix Theory and the SYK Model, Advisor: Antal Jevicki

Publications

○ **Critique-out-Loud Reward Models**
Zachary Ankner*, **Mansheej Paul***, Brandon Cui, Jonathan D. Chang, Prithviraj Ammanabrolu
Under submission. `https://arxiv.org/abs/2408.11791`

○ **Scaling Laws for Precision**
Tanishq Kumar*, Zachary Ankner*, Benjamin F. Spector, Blake Bordelon, Niklas Muennighoff, **Mansheej Paul**, Cengiz Pehlevan, Christopher Re, Aditi Raghunathan
Under submission. `https://arxiv.org/abs/2411.04330`

○ **Does your data spark joy? Performance gains from domain upsampling at the end of training**
Cody Blakeney*, **Mansheej Paul***, Brett W. Larsen*, Sean Owen, Jonathan Frankle
Conference on Language Modeling (COLM), 2024. `https://arxiv.org/abs/2406.03476`

○ **LoRA Learns Less and Forgets Less**
Dan Biderman, Jose Javier Gonzalez Ortiz, Jacob Portes, **Mansheej Paul**, Philip Greengard, Connor Jennings, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, John Patrick Cunningham
Transactions on Machine Learning Research (TMLR), 2024, certified. `https://arxiv.org/abs/2405.09673`

○ **Perplexed by Perplexity: Perplexity-Based Pruning with Small Reference Models**
Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L Leavitt, **Mansheej Paul**
Under submission. `https://arxiv.org/abs/2405.20541`

○ **Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression**
Allan Raventós*, **Mansheej Paul***, Feng Chen, Surya Ganguli
Advances in Neural Information Processing Systems (NeurIPS), 2023.

○ **Unmasking the Lottery Ticket Hypothesis: What's Encoded in a Winning Ticket's Mask?**
Mansheej Paul*, Feng Chen*, Brett W. Larsen*, Jonathan Frankle, Surya Ganguli, Gintare Karolina Dziugaite
International Conference on Learning Representations (ICLR), 2023. **Notable Top 25%**

○ **Lottery Tickets on a Data Diet: Finding Initializations with Sparse Trainable Networks**
Mansheej Paul*, Brett W. Larsen*, Surya Ganguli, Jonathan Frankle, Gintare Karolina Dziugaite
Advances in Neural Information Processing Systems (NeurIPS), 2022

○ **Deep Learning on a Data Diet: Finding Important Examples Early in Training**
Mansheej Paul, Surya Ganguli, Gintare Karolina Dziugaite
Advances in Neural Information Processing Systems 34 (NeurIPS), 2021

○ **Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the Neural Tangent Kernel**
Stanislav Fort*, Gintare Karolina Dziugaite*, **Mansheej Paul**, Sepideh Kharaghani, Daniel M. Roy, Surya Ganguli
Advances in Neural Information Processing Systems 33 (NeurIPS), 2020