Siddharth Verma

https://siddharthverma314.github.io | siddharthverma314@gmail.com

About me

I am an ML Researcher with deep expertise in training state-of-the-art LLMs. I write performant low-level code to train models across thousands of GPUs. I also conduct research in both academic and industrial settings, resulting in publications in prestigious venues like NeurIPS and ACL.

Languages: Python, Haskell, Rust, Java, SQL, C, Go, CUDA, RISC-V
Skills: Language Modeling, LLM Pretraining, Reinforcement Learning, Neural Networks, Statistics
Technologies: Pytorch, JAX, OpenAI Triton, Pallas, Docker, NixOS, Unix/Bash, Git, Google Cloud

Experience

GDM Research Engineer

Google Deepmind | Cambridge MA | Aug 2024-Current

- Core contributor to Gemini Pretraining in both model architecture and implementation
- Wrote a performant pallas kernel for 2-simplical attention to enable research into higher order attention
- Investigated numerical instability in MoE implementation and its effects on pretraining and RL

Research Engineer

Character.ai | New York NY | Dec 2023-Aug 2024

- Contributed to all aspects of LLM pretraining from fundamental research to performant implementations
- Discovered the relation between attention and intelligence in LLMs and its corresponding scaling laws
- Designed scaling laws to predict a 10,000x flops extrapolation in validation loss with tight error bounds
- Implemented multiple MoE variants for our flagship model trained across our entire cluster of GPUs
- Investigated the non-causality of Expert Choice MoEs and when this affects model performance

Senior Machine Learning Engineer

Square | Boston MA | Sep 2022-Dec 2023

- Finetuned open-source LLMs on merchant-buyer conversations to suggest replies to incoming messages
- Conducted an online A/B test and demonstrated a 5% increase in suggestion acceptance rate
- Designed and implemented a multi-task training system to incorporate classification tasks into an LLM
- Instruction finetuned FLAN-T5 on internal data and evaluated performance against individual classifiers

AI Resident

Meta (Facebook) | Seattle WA | Aug 2021-Sep 2022

- Wrote code to process 1TB of multimodal data using Rust and Parquet for a 20x speedup against Python
- Automated the training LLMs of up to 13B parameters on large multi-node clusters with up to 64GPUs
- Evaluated whether training on explanations improve reasoning capabilities of LLMs, and found that explanations mostly benefit mathematical reasoning
- Analyzed effect of masking rates and masking strategies in multimodal learning, showing that increasing masking rate nullifies effects of different masking strategies

Undergraduate Researcher at Robotic AI and Learning Lab

Berkeley Artificial Intelligence Research Lab | Berkeley CA | Jan 2019-May 2021

- Worked with Prof. Sergey Levine and Prof. Chelsea Finn on RL and NLP in domains of robotics and chatbots
- Designed and implemented a multi-agent RL algorithm to learn composable locomotive skills without manual environment resets, subsequently using them to solve a maze. Published at NeurIPS
- Used Offline RL to finetune LLMs to bargain on craigslist items, beating supervised learning in human evals across all metrics. Accepted as oral presentation at NAACL

Education

BA Computer Science & Music

UC Berkeley | GPA: 3.965, Major GPA: 4.0 | Aug 2017-May 2021