PRATIK DOSHI

+1(831) 266-8773 ◇ Santa Cruz, CA
prdoshi@ucsc.edu ◇ linkedin.com/in/pratik-doshi-b2a493153/ ◇ github.com/Pratik-Doshi-99

EDUCATION

MS Computer Science, University of California, Santa Cruz — Expected: Apr 2025
Relevant Coursework: Neural Computation, Deep Learning, Compilers, Linear Algebra — GPA: 3.91/4.0

EXPERIENCE

AI Research Intern — 07/2024 - 09/2024
Data Care LLC, Utah, USA *(LLM Inference, vLLM, Kubernetes, PyTorch, Docker)*

- Developed an LLM throughput analyzer that load tests deployed models and tracks metrics like TTFT, Throughput, and Inter-token latency under heavy concurrent usage conditions at a prompt-level granularity.
- Deployed Open-Source LLMs like Llama-3.1-8B through vLLM using Kubernetes orchestration, and achieved an inference throughput of 700+ tokens/sec on a single NVIDIA L4.
- Researched SOTA inference techniques like ZeRO memory optimizations (DeepSpeed), Flash Attention, Dynamic Batching, and OS engines like vLLM and Triton Inference Server.

Associate Software Engineer — 06/2021 - 03/2023
Rupeeseed Technology Ventures, Mumbai, India *(C#, System Design, Performance Profiling, AWS Cloud Services)*

- Reduced turnaround latency for a recommendation system from 15 minutes to 2 seconds using LINQ in C#.
- Designed a data processing pipeline in C# and improved its throughput by 50% using pipeline parallelism.
- Designed MongoDB schemas and applied Indexing and Sharding strategies to improve read performance and API throughput by more than 90%.
- Developed a statistical model to predict the profitability of customized trading strategies using probability distributions and derivative pricing models.

PROJECTS

Time-series FMs (ongoing) *Pretraining Pipeline, DeepSpeed, Kubernetes, PyTorch, Transformers*
Building a foundation model for off-the-shelf time series prediction. Designed a pretraining pipeline involving synthetic non-stationary time series (1M samples), stock prices (10M samples), and Gaussian mixing.(Github)

Finetuned Code-Llama for Text to SQL task *LLMs, PEFT, LoRA, Huggingface, LLM Evaluations*
Finetuned Code Llama 7B using Parameter Efficient Fine-tuning (PEFT) and the Huggingface library to improve its performance on generating SQL Queries from natural language instructions. (Huggingface)

Image Captioning using VLMs. *Deep Learning, Multi-modal AI, PyTorch, Kubernetes*
Trained a Vision-Language model on the image captioning task and achieved 25% improvement on the BLEU metric, using dynamic attention (from the paper "Show Attend and Tell"). (Github)

Training and Evaluating Sparse Autoencoders *Interpretability, SAEs, JumpReLU, PyTorch, Evaluations*
Trained Sparse Autoencoders using ReLU and JumpReLU activation functions and evaluated them on a feature annotated dataset to assess reconstruction. Achieved MSE below 0.4 with 95% sparsity. (Colab)

Identified a Neural Circuit for a Coding Task *AI Interpretability, LLMs, PyTorch, Google Cloud*
Applied an advanced graph-based algorithm to perform ablations on an LLM's Attention and MLP layers to identify the set of components (neural circuit) responsible for the defined code completion task. (Github)

SKILLS

Machine Learning	LLMs, Inference, Deep Learning, Distributed Training, MLOps
Technical Skills	PyTorch, Python, vLLM, Triton Inference Server, C, C#, REST APIs, MongoDB, SQL
Cloud	AWS EC2, Bedrock; GCP VMs, Kubernetes, Docker, Bash Scripting, CI/CD