PRATIK DOSHI

(831) 266-8773 ◇ Santa Cruz, CA 95060
prdoshi@ucsc.edu ◇ linkedin.com/in/pratik-doshi-b2a493153/ ◇ github.com/Pratik-Doshi-99

SKILLS

Machine Learning	LLMs, VLMs, Inference, Distributed Training, AI Alignment
Technical Skills	PyTorch, Python, vLLM, TensorRT, DeepSpeed, C/C++, C#, REST APIs
Cloud	AWS EC2, Bedrock; GCP VMs, Kubernetes, Docker, Bash Scripting, CI/CD

EXPERIENCE

AI Engineer 03/2025 - Present
Afters, San Francisco Bay Area

- Improved inference latency by 40% with TensorRT compilation for a custom image generation workflow.
- Developed a server to scale inference for Stable Diffusion on top of ComfyUI and TensorRT engines.

AI Research Intern 07/2024 - 09/2024
Data Care LLC, Utah, USA

- Achieved a throughput of 700+ tokens/sec on a single NVIDIA L4 using quantization.
- Developed an LLM throughput analyzer to benchmark LLMs using vLLM.
- Researched SOTA inference techniques Flash Attention, Speculative Decoding, and Paged Attention.

Associate Software Engineer 06/2021 - 03/2023
Rupeeseed Technology Ventures, Mumbai, India

- Developed a statistical model to predict the profitability of trading strategies using derivative pricing models.
- 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.

EDUCATION

MS Computer Science, University of California, Santa Cruz Mar 2025
Relevant Coursework: Neural Computation, Deep Learning, Compilers, Linear Algebra GPA: 3.92/4.0

PROJECTS

Designed a Power-based Hardware Attack *NVIDIA GPU Architecture, CUDA, Deep Learning, Security*
Found a vulnerability that leverages the power draw statistics of NVIDIA GPUs to leak architectural details of the models running on those GPUs. Achieved 90%+ detection accuracy.

Designed a DSL called DataEase *Compilers, Tokenization, Domain Specific Programming Languages*
Designed a grammar and implemented a tokenizer and parser for a domain specific language called DataEase. DataEase makes Data Science easy and compiles to native Python using its compiler. (Github)

Finetuned Code-Llama for Text to SQL task *LLMs, PEFT, LoRA, Huggingface, LLM Evaluations*
Finetuned Code Llama 7B using PEFT to achieve 4% accuracy improvement on generating SQL Queries from natural language instructions. Used Kubernetes to execute the training. (Huggingface)

Implemented a Custom Activation Function *JAX, Neural Network Architectures, Custom Loss Functions*
Implemented a custom activation function and loss metric, trained Sparse Autoencoders using Straight-through Estimators, and improved feature reconstruction by 2% (from the paper on JumpRELU) . (Github)

Improved a VLM with Attention *Deep Learning, VLMs, PyTorch, Kubernetes*
Trained a Vision-Language Model from scratch on the image captioning task and achieved 25% improvement on the BLEU metric, using dynamic attention (from the paper "Show Attend and Tell"). (Github)