PRATIK DOSHI

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

SKILLS

Technical Skills	PyTorch, Python, C/C++, vLLM, Multi-threading, SQL, MongoDB, Redis
Cloud	AWS EC2, Bedrock; GCP VMs, Kubernetes, Docker, Bash Scripting, CI/CD
Finance	Financial Markets, Portfolio Management, Risk Management, Derivatives

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

BS Finance, NMIMS University, Mumbai — Apr 2021
Relevant Coursework: Financial Accounting, Taxation, Portfolio Management — GPA: 3.85/4.0

EXPERIENCE

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

- Achieved an inference 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 PagedAttention.

Backend Engineer, Fintech — 06/2021 - 03/2023
Rupeeseed Technology Ventures, Mumbai, India — *(C#, Multi-threading, Real-time Systems)*

- Reduced turnaround latency for a strategy generation engine from 15 minutes to 2 seconds using LINQ in C#.
- Developed a charting engine and high throughput REST APIs using .NET, MongoDB, MSSQL and Redis. Applied multi-threading in the charting engine to speed up I/O bound tasks and achieve real-time data streaming.
- 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%.

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.

Volatility Prediction in Financial Markets. — *Econometrics, Time Series, Regression Analysis*
Designed a variance prediction model by leveraging the market discrepancies. Successfully reduced average prediction error (statistically significant) of traditional GARCH models by 50%. (Portfolio)

Time-series FMs (ongoing) — *E2E Model Development, Transformers, Deep Learning*
Currently building a transformer-decoder based foundation model for time series prediction. Implementing a custom pretraining pipeline and label smoothing for a robust gradient signal. (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)

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)