Amogh Patankar

www.linkedin.com/in/apatankar22 | github.com/apatankar22 | apatankar22.github.io | San Francisco, CA | (408)-597-2478

Education

University of California, San Diego September 2023 – March 2025
*M.S., Computer Science & Engineering; Concentration: **AI & Machine Learning*** *La Jolla, CA*

- Growth Strategist, **Perplexity AI**
- Co-President, **CSE Graduate Student Council**

University of California, San Diego September 2020 – March 2023
B.S., Data Science *La Jolla, CA*

Professional Experience

Research Engineer May 2025 – Present
Hedra *San Francisco, CA*

- Research and implementation for state-of-the-art mid-training/fine-tuning and inference optimization techniques for an efficient & scalable foundational diffusion model, trained on various distributed multi-node, multi-GPU HPC systems.
- Deployment and product development for Realtime Avatar, a low-latency and style-agnostic realtime avatar model.
- Developing ML agents and agentic evaluation frameworks to reason across text, audio, image, and video, and enabling multi-turn agents using Pydantic AI.
- Responsible for data purchases, ingestion, infrastructure, and orchestrating the deployment of an end-to-end, scalable data pipeline processing hundreds of terabytes of videos daily.

AI Applications Development Intern September 2024 – December 2024
Advanced Micro Devices (AMD) *San Jose, CA*

- Optimized generative AI workload (Llama-2, Llama-3, etc.) execution on RyzenAI neural processing unit (NPU) using caching, batching, quantization, as well as model and data parallelism.
- Benchmarked generative AI workloads on AMD and competitor hardware (Snapdragon X Elite, Lunar Lake); improved Llama-2 throughput by ∼ 5%, with constant time to first token.
- Evaluated NPU latency and utilization during concurrent execution of generative AI workloads and high-fidelity AI effects.

Data Scientist, Generative AI Intern June 2024 – September 2024
Marvell Technology Inc. *Santa Clara, CA*

- Architected an end-to-end ETL pipeline consisting of data engineering, analysis and visualization using numpy, pandas, sklearn, Tableau, etc. Data was ingested using Apache and AWS Kinesis, and stored in SnowflakeDB, and AWS S3.
- Utilized generative pretrained transformer (GPT) models to generate synthetic data, leveraging parameter efficient fine-tuning (PEFT) techniques such as low-rank adaptation (LoRA).
- Implemented reinforcement learning (RL) and deep learning methods for DSP parameter optimization.
- Integrated large language models (LLM) and retrieval augmented generation (RAG) to automate hardware modeling process.

Researcher June 2023 – August 2024
Stanford University School of Medicine *Palo Alto, CA*

- Performed data engineering, analysis, and visualization through various methods in Python and R; specifically, pandas, geopandas, numpy, matplotlib, scikit-learn and ggplot. Developed unique statistical packages composed of chi-squared and Fisher tests.
- Led research teams mentored by Dr. Gross, Dr. Palaniappan, and Jin Long. **Opioid overdose research** published in British Journal of Anaesthesia, and diabetes research in preprint at Journal of Asian Health.

Software Development Engineer Intern June 2022 – September 2022
Amazon Web Services (AWS) *Seattle, WA*

- Led architectural changes in Lex ASR (Automatic Speech Recognition) Services and DataHub, improving latency for conversational AI models. Conducted A/B tests to evaluate architectural changes for customer use.
- Performed cohort analysis to segment critical and non-critical customer data in DataHub, enhacing Lex ASR schemas, and enabled compliant storage of all data using AWS Kinesis, S3, and Lambda.
- Applied time series analysis on AWS CloudWatch metrics to track and optimize the performance of ASR Service, leading to lower response time and accelerating customer request resolution by up to ∼75%.

Skills

- **Languages:** Python, Java, R, C++, SQL
- **Frameworks and Tools:** PyTorch, CUDA, Triton, MPI, Tensorflow, Pydantic, Slurm, ONNX, TogetherAI, Keras, Airflow, numpy, AWS [S3, EC2, SageMaker, Kinesis, Lambda], SnowflakeDB, Tableau, Streamlit, git

Projects

- **LLM Systems Projects**: Autodiff, fusion ops, kernel optimization, Mixture of Experts (MoE), Speculative Decoding
- **Capstone Project**: Active Learning with Neural Processes for Epidemiology Modeling
- Prediction of Causes of Patient Readmission using Large Language Model(s) (**OpenAI, Llama-7B, BART-Large**)
- Data Science Interview Tool (**GPT-3**, PyTorch, Python)