Pranav Vempati

2019 Boren Ave, Seattle, Washington
(408)-507-3418
pranav.k.vempati@gmail.com

Summary

I am a software engineer at AWS working on Just Walk Out (JWO) technology, with a Master's in Computer Science and a background in deep learning research and HPC, including work on generative models and C++ simulation libraries at Lawrence Livermore National Laboratory. I previously served as President of Santa Cruz Artificial Intelligence, a 150+ member ML organization at UC Santa Cruz.

Skills

Programming Languages	Bash, C, C++, Java, JavaScript, Kotlin, Python, SQL, Swift, TypeScript
Frameworks & Libraries	Angular, CUDA, Keras, Matplotlib, NumPy, ONNX, OpenCV, Pandas, PyTorch, Scikit-Learn, SciPy, Seaborn, TensorFlow
Machine Learning & AI	Computer Vision, Data Science, Generative AI, Large Language Models (LLM), Machine Learning, Natural Language Processing, Reinforcement Learning, Transformers
Software Development Tools and Concepts	Agile, Automation, AWS, Big Data, CI/CD, CloudWatch, CMake, Concurrency, DevOps, Distributed Systems, Docker, DynamoDB, Git, GitHub Actions, GPU Acceleration, Jenkins, Kubernetes, Lambda, Linux, Multithreading, Parallel Programming, Performance Analysis, SLURM, Systems Programming

Experience

03/2025 – Present **Software Development Engineer - Just Walk Out (AWS)**, *Amazon*, Seattle, Washington

08/2022 - 01/2023 **Data Scientist**, *Lawrence Livermore National Laboratory*, Livermore, California

- Implemented Continuous Conditional GANs (CcGANs) in PyTorch as a learned surrogate for spatial laser energy deposition during 3D metal printing, replacing classical melt pool simulations that require HPC-scale compute to resolve turbulence and shock effects in molten metal.
- Modified the CcGAN loss function to incorporate physical constraints of the deposition process and built custom dataset augmentation pipelines for image data; achieved 93% fidelity to ground truth physical experiments.
- Worked on libROM, a C++ ROM (Reduced Order Modeling) library to reduce the computational overhead of physical simulations. Authored a comprehensive suite of regression tests, enhanced the library's continuous integration workflow using GitHub Actions, and incorporated Finite Element Method based simulations into libROM.

06/2024 – 03/2025 **Machine Learning Engineer**, *CoreData AI*, Remote
- Engineered and deployed several AI chatbots and backend models enabling customers to derive actionable insights from their data, improving time-to-decision by 60% and engagement by 40%.
- Developed the backend in Python on AWS Lambda, exposing scalable endpoints via AWS API Gateway for low-latency, serverless inference.
- Developed domain-specific prompt engineering and response templates (task decomposition, structured outputs, guardrails) tailored to customer workflows, improving answer relevance, consistency, and reducing hallucinations across common query patterns.

09/2021 – 08/2022 **President**, *Santa Cruz Artificial Intelligence*, Santa Cruz, California
- Led a UCSC Baskin School of Engineering affiliated organization comprising over 150 members.
- Prepared and delivered lessons to members, and mentored members as they worked on their projects.

09/2018 – 09/2021 **Lecturer**, *Santa Cruz Artificial Intelligence*, Santa Cruz, California
- Delivered weekly lessons to members in conjunction with other officers.

06/2021 – 06/2022 **Researcher**, *UCSC Computer Vision Lab*, Santa Cruz, California
- Worked with Professor Roberto Manduchi to maintain and benchmark an iOS Swift-based OCR application for the visually impaired.
- Incorporated features of the Google MLKit iOS API into the application.

06/2020 - 09/2020 **Software Engineering Intern - End User Computing**, *VMware*, Palo Alto, California
- Implemented an Angular-based project to enhance the user experience for the renovated Workspace One administration console.
- Added new functionality to the Identity Management UI in Angular 8 and Clarity.

08/2019 - 09/2019 **Embedded Systems and Deep Learning Intern**, *ICURO*, Santa Clara, California
- Delivered a Python OCR-based Computer Vision license plate recognition system leveraging TensorFlow Lite-based MobileNet and Darknet YOLO v3 models.
- Achieved 96% accuracy on a custom-scraped dataset on resource-constrained Edge TPU hardware.
- Delivered 8-bit quantized models for on-device offline inference on Mendel Linux embedded systems, realizing a 10x reduction in model size.

Education

01/2023 - 06/2024 **M.S. in Computer Science**, *University of California, Santa Cruz*, Santa Cruz, California, GPA: 3.82/4.00

Relevant Coursework Computer Vision, Artificial Intelligence, Research and Teaching in Computer Science, Applied Bayesian Statistics, Analysis of Algorithms, Computer Architecture, Compiler Design, Introduction of Probability Theory

MS Project *Evaluating the Effectiveness of Fairness Techniques for Decision Tree Classifiers*. Implemented and benchmarked five fairness interventions including rule-based pruning, Shapley-importance pruning, and direct modification of Scikit-learn's Cython Decision Tree training algorithm to penalize imbalanced splits; the latter approach reduced false positive rate divergence on the maximally divergent subgroup by 7.82% with no degradation in test accuracy or ROC AUC.

09/2018 - 12/2021 **B.S. in Computer Science**, *University of California, Santa Cruz*, Santa Cruz, California, GPA: 3.46/4.00