Pags

☎ 469-426-9880 San Francisco, CA ✉ ananyapag@gmail.com /ananyapg /ananyapags pags.dev

Software Engineer building scalable AI infrastructure and HPC systems, and then making them faster just for fun.

AI/ML Tools: PyTorch, LangChain, Weaviate, Pinecone, vLLM, Metaflow, CUDA, Triton, OpenMP, MPI, Apache Spark
Languages: Python, C++, Go
Developer Tools: AWS, GCP, Docker, Kubernetes, Slurm, Linux, React, Node.js, Flutter, Git, GitHub Actions, Figma

Education

Santa Clara University

Masters of Computer Science and Engineering, Systems and Performance Emphasis

- Research: Evaluation of Optimization Techniques for Large Language Model Inference
- *Relevant Coursework: Computer Architecture, Cybersecurity, Distributed Systems, OS, High Performance Networks*

Bachelors of Computer Science, Cybersecurity Emphasis

- Spanish Language and Mathematics minor; Critical Thinking and Writing 1 & 2 Teaching Assistant
- Research: Implementing and benchmarking facial recognition pipelines in public safety instances

Projects

LABUBU AI Shopper *(MCP, GPT-4, Python, FastAPI, React, Firebase, Faiss)*

- Built a POPMART **shopping bot** monitoring **1,000+ SKUs** with **200ms response time**, **95% match accuracy**, and **sub-second automated checkout** using vector search.

GPU-Accelerated Image Processing Pipeline *(CUDA, PyTorch, C++, NVIDIA GPU)*

- Designed **GPU-accelerated image filters** (blur, sharpen) benchmarked on **NVIDIA V100 GPUs**, achieving **8× speedup** over CPU baselines with PyTorch preprocessing.

High-Performance Matrix Inversion via Custom GPU Kernels *(Triton, OpenCL, C++)*

- Implemented **custom GPU kernels** with **Triton** for **5,000×5,000 matrices**, achieving **50% runtime reduction** vs CPU inversion using parallelized **OpenCL pipelines**

University Esports Discord Chat Content Moderation System *(AWS Bedrock, Python, Lambda, SageMaker, DynamoDB)*

- Deployed on **7 Discord servers**, classifying **8k+ messages/min** with **96% accuracy** and **<100ms latency** using AWS Lambda.

Parking Ticket Prevention Software *(Spark, Kafka, Lambda)*

- Designed a **real-time streaming platform** to track **parking enforcement vehicles**, sending **live alerts via Kafka event streams** for proactive compliance.

Experience

Support Engineer, WaveHPC — May 2025 - Now

- Built and maintained **scalable HPC workflows** supporting **multi-GPU training** and **large-scale benchmarking**.
- Designed **automation pipelines** for onboarding, job scheduling, and workload monitoring using **Slurm + Kubernetes**.
- Improved **cluster utilization by 18%** through profiling, load balancing, and resource optimization.

Software Engineer, Frugal Innovation Hub — Jun 2024 - May 2025

- Developed a **bilingual math learning platform** (**Flutter, Firebase, Firestore, OAuth2**) for **200+ active users**.
- Integrated **real-time synchronization** and **secure authentication**, improving performance by **23%**.
- Implemented **CI/CD** with **GitHub Actions + Docker**, streamlining release cycles and reducing deployment time by **40%**.

Machine Learning Research Engineer, National Science Foundation — Jun 2023 - Sep 2023

- Developed **synthetic malware** simulators using **PyTorch** to benchmark **detection efficiency** on dynamic threat datasets.
- **Automated ETL processes** with preprocessing, dimensionality reduction, and feature selection on Malicia's **50+ datasets**.
- Addressed **class imbalance** using **GANs, SMOTE, ADASYN**, improving detection accuracy by **17%**.

Forward Deployed Engineer Miller Center for Social Entrepreneurship x SuitUp — Jan 2023 - Nov 2023

- Enabled **data-driven scaling decisions** for Fortune 500 and nonprofit partners with **1,000+ qualitative data points** analyzed.
- Trained **DistilBERT NLP models** for automated **text classification** of qualitative interviews.
- Automated **Salesforce CRM workflows**, cutting manual data entry time by **12%**.

President, ACM-W (Women's Computer Science Club @ Santa Clara University) — May 2022 - Jun 2023

- Directed **technical operations** for 12 SWEs, delivering **event registration & applicant management systems** for **500+ users**.
- Organized **8 workshops**, **2 summits**, and **4 hackathons** (**400+ attendees each**), driving regional tech engagement.

Computer Science Instructor, Juni Learning — Dec 2020 - Mar 2023

- Taught **Python (ML, gaming)** & **C++ (competitive programming)** to **45+ students**.
- Created **200+ tailored lesson plans & projects**, adapting pedagogy to student goals and progress.

Awards

- Publication @Santa Clara University 2024 - "SuitUp: Addressing Employee Retention, Satisfaction, and Scaling"
- Publication @EAI Intetain 2020 - One of 10 posters selected to present "Fixing AI for Public Safety"
- WON: Most Interdisciplinary Award at Hack For Humanity 2021: *Devpost:* locals-n9r2u3
- WON: Fourth place at Bronco CTF (Capture The Flag) in Santa Clara, CA