# Pags

469-426-9880 San Jose, CA apagadala@scu.edu /ananyapg /ananyapags pags.dev

Passionate about building agentic AI systems and optimize Large Language Model performance using High Performance Computing, combining cutting-edge technologies with scalable, real-world engineering.

**AI/ML Tools**: PyTorch, BERT, OPT-125, GPT-4, DeepSeek, LangChain, Weaviate, Pinecone, Astra DB, CUDA, Triton, OpenMP
**Languages**: Python, C++, Git, Go, Rust, Assembly
**Developer Tools**: AWS, GCP, Docker, Kubernetes, Metaflow, Linux, React, Flutter, Figma, Langflow, Twilio

## 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 (current), Operating Systems (current)*

Bachelors of **Computer Science**, Cybersecurity Emphasis

- Spanish Language and Mathematics minor; Critical Thinking and Writing 1 & 2 Teaching Assistant
- Research: Exploring symmetries in Graph Coloring Reconfigurations
- Twitch Affiliate 2022

## Selected Projects

**Sales Bot *(GPT-4, Pinecone, Python, FastAPI, React, Firebase, OAuth2, Faiss)***

- LLM-powered sales chatbot using vector stores to enable sales representatives to make queries using natural language

**AI-powered Content Moderation System with AWS Bedrock *(AWS Bedrock, Python, Lambda, SageMaker, DynamoDB)***

- Built an AI content moderation system using AWS Bedrock for real-time text classification, integrated with Lambda and DynamoDB, achieving 95% accuracy and reducing manual moderation by 60%

**Real-Time Fraud Detection System for E-commerce *(Scala, Spark, Kafka, Lambda)***

- Developed a real-time fraud detection pipeline using Spark and Kafka, deploying on AWS Lambda for serverless, low-latency processing of streaming transaction data, achieving a sub-100ms response time and 98% fraud detection precision

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

- A GPU-accelerated project that implements image filters (blur, sharpen) using CUDA for parallel processing, with PyTorch for image handling and performance comparison to CPU

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

- Developed an optimized matrix inversion algorithm using Triton for GPU kernel generation and OpenCL for cross-platform parallel processing, reducing computational time by 50% for large-scale matrices (up to 5,000x5,000 elements)

## Experience

**Software Engineer,** Frugal Innovation Hub — Jun 2024 - now

- Engineered a scalable, **agentic based** bilingual mathematics **mobile** learning platform for **200+ active users** using Flutter
- Integrated Firestore and OAuth2 for **real-time sync** and **secure authentication**, improving performance by **23%**
- Developed a **responsive UI** with **i18n** for localization and **WCAG 2.1** accessibility, increasing engagement by **40%**

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

- Designed **scalable ML workflows** using Pytorch and TensorFlow for malware detection in dynamic threat environments
- **Automated ETL processes** with preprocessing, dimensionality reduction, and feature selection on Malicia's **50+ datasets**
- Used **synthetic oversampling** (GANs, S.M.O.T.E., ADASYN) to handle class imbalance, improving model performance
- Applied hybrid LSTM-Transformer models, achieving **98%** accuracy via **Bayesian hyperparameter tuning and validation**

**Machine Learning Research Fellow,** Miller Center for Social Entrepreneurship x SuitUp — Jan 2023 - Nov 2023

- Selected as 1 of 16 students awarded the 2023 Miller Center Lewis Family **Research Fellowship and Scholarship**
- Used **NLP** to train **DistilBERT** for **automated text classification** of qualitative interviews, for data-driven scaling decisions
- Streamlined CRM in **Salesforce** through **real-time data organization** for a **12%** reduction in working capital

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

- Promoted tech literacy through **8 workshops, two summits, and 4 Hackathons** for audiences across the Bay Area
- Co-directed Hack for Humanity, Santa Clara University's **largest** student-run Hackathon with **400+ attendees**
- Supervised 12 developers to build dynamic registration and **applicant management systems** for **500+users**

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

- Taught **45+ clients** Python (**ML and Gaming**) and C++ **(Competitive Programming)** curriculum to achieve long term goals
- Customized tutoring and instructional approaches by writing tailored lesson plans with **200+ unique projects**
- Incorporated clients strengths and weaknesses into sessions, while maintaining a log of growth with **1000+ entries**

## 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