Pablo Leyva

908-525-6062 | pl33@njit.edu | linkedin.com/in/pablo-leyva | github.com/pleyva2004 | pabloleyva.io

Education

New Jersey Institute of Technology — Newark, NJ

B.S. in Applied Mathematics (Statistics) & Computer Science (AI Concentration) — *Aug 2023 – May 2027*

Relevant Coursework: Data Structures & Algorithms, Computer Vision, Linear Algebra, Probability & Statistics, Database Systems, Software Engineering, Calculus III, Numerical Methods, Objected Oriented Programming, **Deep Learning PhD Coursework**

Experience

Apple — May 2025 – August 2025

AI/ML Product Engineering Intern — *Cupertino, CA*

- Architected **multi-agent orchestration system** integrating the Apple Foundation Model(AFM) with Model Context Protocol (MCP), implementing **graphRAG** and **Source of Truth** algorithm to ground every answer with a verifiable source
- Designed **tool selection, execution, and deployment** with Basemind(LangGraph) which used **TouchID as Human in the Loop validation** for deterministic agent behavior in the Apple Pay Agentic Payment flow
- Presented the prototype to higher leadership to show the feasiblity of the **Agentic Shopping experience** and the potential of the Apple Foundation Model(AFM) for the future of Apple Pay

Undergrad Research and Innovation (URI) - Agentic Reasoning SLM — Sep 2025 – Present

Research Fellow — *Newark, NJ*

- Objective: Build the **specialized reasoning brain** of an agentic system, and let conversational models handle the user interaction
- Training a 3B parameter SLM architecture that is trained for one-step reasoning and tool selection
- Constructing **mathematical logic and proofs dataset** as well as **logic puzzles dataset** to train the SLM
- Evaluating the agentic reliability of the model in executing long, deterministic step-by-step tasks vs. larger LLM baselines
- Integrating **lightweight adapter layers (LoRA)** and **Reinforcement Learning** into the model to specialize it for structured business workflows

Caterpillar Inc. — May 2024 – August 2024

Software Engineering Intern — *Chicago, IL*

- Architected **end-to-end data pipeline** ingesting 500K+ daily records from Slack, Jira, and Confluence APIs, building ETL workflows with Python and Apache Airflow to power ML models for team productivity analytics
- Fine-tuned **transformer-based NLP models** (BERT, GPT-3.5) to extract semantic insights from unstructured communication data, achieving 89% accuracy in classifying collaboration patterns and identifying project blockers
- Engineered **RESTful microservices** with FastAPI to serve model predictions in production, handling 10K+ daily inference requests with sub-100ms latency using Redis caching and async processing
- Built **interactive React dashboard** with D3.js visualizations, enabling leadership to explore predictive analytics on team velocity and resource allocation optimization

Projects

Open Source - NeuroCache | *ONNX, Transformers.js, WebAssembly, PyTorch, llama_cpp* | **Python**

- Implemented **custom KV-cache manipulation layer** with **dynamic context compression** via attention-based importance scoring and memory slot management, in the proces of benchmarking memory efficiency and response quality across conversation sessions

Multilingual Voice Agent – Inventory Ordering System | *FastAPI, WebSockets, Whisper, gTTS, OpenAI API* | **Python, JavaScript**

- Developed **real-time multilingual voice agent** supporting Spanish, Portuguese, Arabic and English for mechanic shop workers, implementing **speech-to-text transcription** with Whisper API and **context-aware LLM processing** to autonomously place inventory orders based on natural voice conversations

Emotion-Aware Music Recommendation System | *PyTorch, BERT, librosa, Spotify API, Flask* | **Python**

- Architected **multi-modal deep learning pipeline** with separate CNN branches for mel-spectrogram analysis and BERT transformers for lyrics sentiment processing, applied **domain adaptation techniques** for music recommendation and developed **evaluation metrics** to benchmark emotion classification accuracy

Community Outreach

President, Vice President - External Affairs, Event Coordinator — October 2023 – Present

Society of Hispanic Professional Engineers (SHPE) — *Newark, NJ*

- Confidently leading a team of 20 to organize professional events for 300+ students across campus using articulate communication
- Grew our External Company connectiosn by 6x in only two months through confident outreach and relationship building
- Successfully breaking attendance records on a weekly basis by averaging 200% growth in member engagement

Technical Skills

Languages : Python, C++, Java

Frameworks : PyTorch, TensorFlow, scikit-learn, LangChain, LangGraph, MCP

Libraries : Jupyter Notebook, Pandas, NumPy, Pytest, scipy, FastAPI, Flask

Machine Learning : Linear Regression, Logistic Regression, K-Means Clustering, K-Nearest Neighbors (KNN)

AI/ML : Natural Language Processing (NLP), Neural Networks (DNN, CNN, RNN), Deep Learning, Reinforcement Learning

MLOps : Model Deployment, Data Pipelines, Docker, CI/CD, AWS, GCP, Azure, GitHub Actions, linux, kubernetes, unix