

# Pablo Leyva

908-525-6062 | [pl33@njit.edu](mailto:pl33@njit.edu) | [linkedin.com/in/pablo-leyva](https://www.linkedin.com/in/pablo-leyva) | [github.com/pleyva2004](https://github.com/pleyva2004) | [pabloleyva.io](https://pabloleyva.io)

## EDUCATION

### New Jersey Institute of Technology

B.S. in Applied Mathematics ( Statistics ) & Computer Science (AI Concentration)

Newark, NJ

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

### Undergrad Researcher (CAHSI) - Multilingual LLM Training and Bias Analysis

October 2025 – Present

Research Fellow

Newark, NJ

- Build and train a **foundational multilingual model** on Spanish, Portuguese, Italian, Farsi, and Arabic, with the primary objective of enabling the model to "think" and reason in native languages rather than through translation
- Compare and analyze **bias patterns** between models trained exclusively in English versus multilingual models trained on diverse language corpora to evaluate cultural and linguistic bias disparities
- Train specialized **Small Language Models (SLMs)** on individual target languages and benchmark performance against a foundational multilingual LLM to **compare efficiency** versus **accuracy tradeoffs**
- To evaluate model performance across multilingual tasks, measuring both computational efficiency and linguistic fidelity in native language reasoning

### Undergrad Research and Innovation (URI) - Efficient SLM for Agentic Workflows

Sep 2025 – Present

Research Fellow

Newark, NJ

- Designing and building 3B parameter SLM architecture that is trained for CoT and tool selection reasoning
- Integrate lightweight adapter layers (LoRA) and Reinforcement Learning into specialize the model for structured business workflows
- Evaluate the agentic reliability of the model in executing long, deterministic step-by-step tasks vs. larger LLM baselines
- Benchmark on-device performance to **improve performance** and demonstrate feasibility on consumer hardware
- The objective is for the model's foundation to be CoT and step-by-step reasoning so that it can go through post-training process in different niches and carry over that foundation into the specific business use case

### Apple

May 2025 – August 2025

AI/ML Product Engineering Intern

Cupertino, CA

- Architected **multi-agent orchestration system** integrating LLMs with Model Context Protocol (MCP), implementing **reasoning frameworks** and agent coordination patterns using TypeScript and Python
- Designed **agentic workflow pipelines** with LangChain & LangGraph, implementing **prompt flow automation** and validation frameworks to ensure deterministic agent behavior across diverse execution paths
- Led **benchmark development and A/B testing framework** to improve business insights, with a focus on Partner Facing and internal use cases

## PROJECTS

### NeuroCache | WebGPU, 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, developed **performance metrics** to benchmark 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 connections 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++ Typescript, JavaScript, Java

**Frameworks :** React, NEXT.JS , Node, 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