

# 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

## TECHNICAL SKILLS

**Languages :** Python, Typescript, JavaScript, Java, C++

**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

## EXPERIENCE

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

October 2025 – Present

*Research Candidate*

*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** (latency, memory, energy consumption) versus **accuracy tradeoffs** between SLMs and LLMs
- Apply **research techniques** and **algorithm design** 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 Candidate*

*Newark, NJ*

- Apply **research techniques** and **algorithm** design to develop an AFM-style 3B parameter SLM architecture compatible with BitNet b1.58 ternary weights
- Integrate lightweight adapter layers (e.g., LoRA) to 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 (latency, memory, energy) to **improve performance** and demonstrate feasibility on consumer hardware

### Apple

May 2025 – August 2025

*AIML Product Engineering Intern*

*Cupertino, CA*

- Led a team of 3 interns to build MVP for Agentic Payment flow experience for higher leadership at Apple Pay
- Prototyped agent workflows combining **Large Language Model(LLM)-based product** recommendations with Apple Pay checkout experiences, leveraging Typescript to implement Model Context Protocol (MCP) for merchant catalog parsing and transactional automation
- Used **LangChain & LangGraph** to facilitate agent infrastructure and validation
- Led the design and evaluation of a Graph-RAG architecture to enhance knowledge discovery and information relevance across partner use cases, focused on **Partner Facing RAG ChatBot** to improve business insights and decision making

## PROJECTS

### Princeton Hackathon - NeuroCache | WebGPU, Transformers.js, WebAssembly, PyTorch, CoreML / MLX | Python

- Architected **browser-based inference system** using Web LLM and WebAssembly to deploy quantized SLMs (Phi-3.5-mini, TinyLlama) for on-device computation, eliminating server dependencies and achieving sub-second latency through WebGPU acceleration
- Implemented **custom KV-cache manipulation layer** enabling selective memory retention/eviction, dynamic context compression via attention-based importance scoring, and memory slot management for persistent information storage across conversation sessions
- Designed **advanced context control mechanisms** including cache-level RAG injection, custom attention patterns (sliding window, sparse attention), and dynamic cache zone management for fine-grained conversation context control and privacy-preserving forget operations

### GroupGPT – Collaborative Ideation Platform | Next.js, Supabase, Socket.IO, OpenAI API | Python, JavaScript

- Built and prototyped a collaborative chat interface with Next.js and Supabase Realtime, allowing multiple users to contribute simultaneously to a persistent conversation thread
- Integrated OpenAI's GPT models with conversation memory to provide context-aware summarization, clustering of ideas, and automated agenda generation
- Explored agentic design patterns to support role-specific AI assistants (e.g., Sponsorship Agent, Logistics Agent), showcasing use cases in **real-time conversational AI, context-aware agent workflows, and deploying Large-driven features in user-facing apps**

### 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, BERT transformers for lyrics sentiment processing, and embedding layers for metadata fusion
- Implemented **late-fusion neural architecture** combining audio, text, and metadata features through fully-connected layers, achieving emotion/genre classification for sequential recommendation modeling using Transformer encoders
- Deployed end-to-end system with **Seq2Seq prediction model** for next-song recommendations, integrated with Spotify, Last.fm, and Genius APIs, featuring real-time mood trajectory visualization and personalized playlist generation

### Co-Founder

Sep 2024 – Present

*Association of Latino Professionals For America (ALPFA)*

*Newark, NJ*

- Collaborated confidently to start the first professional latino organization on campus for non-engineering students, demonstrating ability to work independently
- Started the external partners affairs network with local ALPFA chapters and companies through articulate networking and intuitive relationship building

### President

March 2025 – Present

*Society of Hispanic Professional Engineers (SHPE)*

*Newark, NJ*

- Built CLARA - AI Assistant to help with organization operations including email management, event planning, outreach, marketing and everything in between, working independently to deliver intuitive solutions
- Confidently leading team of 20 to organize professional development events for a memberbase of 300 + across campus using articulate communication
- Articulated strategic vision to 2x the number of students getting interviews and internships for the latino community through confident mentorship
- Raised \$70,000 capital for the organization to take 100 students to the national conference and operate independently of the university

### Vice President - External Affairs

May 2024 – March 2025

*Society of Hispanic Professional Engineers (SHPE)*

*Newark, NJ*

- Grew our External Organizations and Company connection by 6x in only two months through confident outreach, working independently to identify partnerships
- Currently leading and managing a team of 13 to organize 60+ events and social events a semester to bring social and professional enrichment to our member base of 300+ students using intuitive planning
- Successfully breaking attendance records on a weekly basis by averaging 200% growth in member engagement through articulate event promotion
- Started Webmaster position for NJIT SHPE yielding 81% growth in CS student participation
- Started SHPEtina position for NJIT SHPE yielding 43% growth in Latina student participation through intuitive programming
- Initiated mentorship program where up to 20+ students learned from the experience of the Executive Board
- Established relationships with exterior clubs from local universities to unite the Latino community in NJ/NYC area

### Event Coordinator

Oct 2024 – May 2024

*Society of Hispanic Professional Engineers (SHPE)*

*Newark, NJ*

- Organize commodities such as food and workshops resulting in an average of 65 + student involvement, working independently on logistics
- Collaboratively work with a team of 12 to successfully leading SHPE to be the largest student organization on campus
- Run weekly events yielding NJIT SHPE to receive many accolades and be recognized by the Governor of Newark