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**

Technical Skills

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

ML Frameworks: PyTorch, TensorFlow, JAX (experience), HuggingFace Transformers, Lightning

Audio & Speech: Whisper, Librosa, FFmpeg, spectrogram processing, mel-scale features, streaming ASR pipelines

LLM + Inference: llama.cpp (custom builds), KV-cache optimization, quantization (int4/int8, FP16/BF16), on-device inference, LoRA/adapter tuning, model serving (FastAPI/Flask), LangChain, LangGraph, MCP

MLOps: Docker, GitHub Actions, CI/CD, Linux/Unix, Kubernetes, AWS/GCP/Azure, model deployment & data pipelines

Experience

Undergraduate Researcher (CAHSI) – Multilingual LLM Training and Bias Analysis October 2025 – Present
Research Fellow *Newark, NJ*

- Developing multilingual language models trained across Spanish, Portuguese, Italian, Farsi, and Arabic to evaluate how language-specific patterns influence comprehension, reasoning, and user interaction quality
- Investigating how linguistic structure and cultural context affect model behavior by comparing mono-language and multilingual models across real-world communication tasks
- Training compact models specialized in individual languages to understand performance trade-offs in real-time conversational applications requiring low-latency processing
- Building evaluation suites to measure consistency, stability, and semantic fidelity of responses across languages and accents in dynamic dialogue settings
- Evaluated model behavior across accents, code-switching, and varied prosodic patterns in real conversational contexts

Undergraduate Research & Innovation (URI) – Efficient SLMs for Agentic Workflows Sep 2025 – Present
Research Fellow *Newark, NJ*

- Designing a 3B-parameter model optimized for stepwise reasoning and interactive task execution, supporting rapid back-and-forth exchanges with users and tools
- Integrating lightweight adapters and reinforcement signals to enable the model to follow structured instructions, maintain contextual continuity, and adapt to dynamic inputs
- Studying reliability of model behavior in long, branching multi-turn tasks that mirror real conversational agent scenarios used in enterprise voice assistants
- Benchmarking responsiveness, output coherence, and interaction stability on consumer hardware to evaluate feasibility for always-on agent systems

Apple May 2025 – August 2025
AI/ML Product Engineering Intern *Cupertino, CA*

- Designed and optimized a multimodal voice agent using Apple's AFM models, improving real-time responsiveness by applying asynchronous agent architecture
- Designed task-oriented workflow pipelines using TypeScript, Python, LangGraph and MCP focusing on predictable behavior during rapid user-model interactions
- Developed validation frameworks to ensure consistency, clarity, and user-friendly model outputs across a variety of dynamic prompts and enterprise-facing scenarios
- Collaborated cross-functionally to test model behavior under diverse conversational styles, ensuring smooth transitions between reasoning, action, and response generation
- Focused on reducing conversational latency and smoothing interaction flow under rapid user input

Projects

NeuroCache - Efficient Inference Caching | *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

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**

- Built a **multi-modal emotion recognition pipeline** combining mel-spectrogram CNN features, BERT-based lyric sentiment, and **speaker embeddings** to capture expressive vocal cues from user singing
- Integrated ECAPA-TDNN speaker encoders to model **prosody, intensity, and vocal affect**, enabling personalized recommendations based on a user's emotional expression rather than only audio content