GERONIMO MORALEZ IV

Staff Machine Learning Engineer · AI Systems Architect · Technical Director

Los Angeles, CA | gmora7@gmail.com | gmiv.io | github.com/gmiv | linkedin.com/in/gmfour

EXECUTIVE SUMMARY

Staff ML engineer and systems architect with 18+ years spanning machine learning, VFX, XR, real-time graphics, and voice synthesis. Designs and operates production AI systems end to end—sovereign GPU inference on in-house Blackwell hardware, multi-agent orchestration, LLM safety engineering, and cloud-native product delivery—including a live consumer AI product and paid enterprise modernization work. Known for adversarial self-audit discipline: systems ship with measured claims and verified outcomes, not vibes.

SIGNATURE STRENGTHS

- Multi-agent AI systems: orchestration runtimes, coordination infrastructure, verified production pipelines
- Sovereign LLM infrastructure: trillion-parameter-class MoE serving, INT4 quantization, vLLM, unit-economics analysis
- Semantic search, knowledge graphs, and graph-RAG (Neo4j, Qdrant, multimodal embeddings)
- LLM reliability and safety engineering: error-corrected execution, verification schemes, statistical certification, crisis-detection pipelines
- Voice synthesis, TTS optimization, and audio intelligence pipelines
- Real-time graphics, physics simulation, and spatial computing
- Rapid prototyping to production in high-ambiguity environments

TECHNICAL EXPERTISE

AI / Machine Learning — Multi-agent orchestration (custom frameworks, AutoGen) · LLM deployment and inference optimization (vLLM, INT4/AWQ quantized MoE serving, speculative decoding, routing gateways) · Graph neural networks (PyTorch Geometric, GAT) · Graph-RAG and knowledge graphs (Neo4j) · Semantic search and multimodal embeddings (CLIP, vision-language models) · Vector databases (Qdrant, Pinecone, FAISS) · Voice synthesis and TTS (prosody optimization, neural vocoders) · Object detection and perception (YOLO) · Synthetic data generation and corpus engineering · Model evaluation, alignment, and LLM safety systems

Infrastructure & Compute — High-performance GPU systems (NVIDIA B200 Blackwell, RTX A6000, multi-GPU clusters) · Triton kernel optimization · AWS (Lambda, CDK, DynamoDB, Cognito, CloudFront, WAF, ECS, S3, Bedrock) · Terraform, Docker, Kubernetes · CI/CD (GitHub Actions) · Redis Streams messaging · Scalable backend and API design

Graphics, VFX & Real-Time — Unity, Unreal Engine, XR/AR Foundation · Shader development (HLSL, Metal) · Physics simulation and fluid dynamics · Color science (OKLAB, film sensitometry, 3D LUTs) · Nuke, Houdini, Maya, Blender pipelines · Procedural content generation · Real-time visualization and digital twins

Languages — Python · C# · C++ · Rust · Dart · JavaScript/TypeScript · Swift · Shell/Bash

PROFESSIONAL EXPERIENCE

VFX Pipeline & Machine Learning Engineer | Crafty Apes VFX (Remote) — Oct 2024–Present

- Designed semantic asset-discovery systems for VFX workflows using multimodal embeddings, visual feature extraction, metadata enrichment, and vector similarity search
- Developed artist-aligned taxonomies and ontology layers translating concepts such as motion, texture, composition, and visual energy into machine-readable retrieval signals
- Built indexing, ranking, and evaluation pipelines enabling fast, meaning-based search across large visual-asset collections
- Prototyped AI-assisted workflow integrations connecting machine-learning capabilities with established VFX tools and production pipelines
- Collaborated with creative and technical stakeholders to translate ambiguous production needs into testable ML systems and practical artist-facing tools

Founder & Principal Technologist | Nimo Labs LLC (Los Angeles, CA) — 2017–Present

Boutique R&D studio delivering AI systems, voice synthesis, XR applications, and cloud-native platforms for enterprise and startup clients—alongside a proprietary internal R&D portfolio in agentic engineering systems, applied ML, and consumer AI (specifics available in conversation, under NDA where required).

- Sovereign LLM infrastructure: deploy and operate trillion-parameter-class MoE inference (INT4/AWQ quantization, vLLM) on in-house 4× NVIDIA B200 Blackwell and RTX A6000 hardware—benchmark-driven serving optimization with measured 1,100+ tok/s sustained throughput and self-hosted unit-economics analysis

- Multi-agent production systems: design, build, and operate agentic engineering infrastructure end to end—orchestration runtimes, model-routing gateways, coordination messaging, automated verification with evidence-anchored audit trails, and human-in-the-loop approval gates—used daily to ship real products
- Live consumer AI product: built and operate an LLM-powered companion app in production on AWS serverless (17 CDK stacks—Cognito, DynamoDB, CloudFront, WAF, SES) with a defense-in-depth safety stack: fail-closed age gating, multilingual crisis detection with automated production canaries, jailbreak firewall, moderation and abuse ladders; Flutter client with Cognito SRP auth, Google SSO (PKCE), and Google Play Billing with server-authoritative entitlements
- Applied ML research: graph neural networks for creative-intelligence applications (PyTorch Geometric, CLIP joint embedding spaces, perceptual color science); large-scale training-corpus engineering with adversarial leakage auditing; LLM reliability research—error-corrected long-horizon execution with statistically certified reliability bounds
- Client delivery—legacy platform modernization (2026): replaced an ASP.NET 3.5 WebForms (VB.NET) and SQL Server platform with a React/TypeScript SPA on AWS serverless (37 Lambdas, 33 routes) for a San Diego legal-services firm—fixed-fee milestone engagement, delivered and paid in full, carrying a HIPAA/PHI workstream; earned a scoped autonomous-deploy authorization in production
- Voice/TTS consulting: production voice synthesis systems—prosody optimization, neural vocoder integration, and iOS native speech framework implementations ($25K+ engagements); measured 30–50% TTS inference speedups via FP16 and KV-cache optimization
- A.R.I.E.S.: multi-agent autonomous research system performing literature review, synthesis, and knowledge extraction

Tech Lead — XR Learning Management System & Forestry Analytics | Analog Digital Systems (Remote) — Sept 2024–May 2025

- Virtual Twin modeling and simulation of forest environments
- Sole architect of an enterprise XR-based LMS integrating Unity XR, computer vision, real-time analytics, and AI-driven assessment automation
- Built immersive training simulations with object recognition, sensor fusion, and adaptive difficulty scaling based on learner performance
- Designed offline-capable field analytics tools for harsh, low-connectivity environments—critical for forestry operations
- Implemented cloud backends, telemetry pipelines, and data-visualization dashboards for scalable multi-site deployment

Tech Lead — AR Music Platform & Web3 Systems | Immuse (Remote) — May 2022–Oct 2023

- Led engineering for an AR music platform integrating spatial computing, blockchain infrastructure, and real-time performance capture
- Developed NFT minting pipelines, smart-contract tooling, and Web3 asset-management infrastructure
- Built native iOS plugins for AR positioning, spatial audio, and immersive fan-engagement experiences
- Directed cross-functional teams spanning product, creative, and engineering through multiple release cycles

Real-Time Pipeline Developer — Special Projects Group | Apple (Cupertino, CA) — Mar 2021–Oct 2021

- Conducted R&D on next-generation AR material systems, shader optimization, and real-time rendering pipelines under strict confidentiality
- Developed physics-based water simulations for Apple Watch keynote
- Collaborated with research and product teams on advanced spatial-computing workflows and visual-fidelity benchmarks

Senior Pipeline Developer — Episodic VFX | Encore Hollywood (Hollywood, CA) — Jul 2017–Aug 2018

- Managed and scaled VFX pipelines supporting thousands of episodic shots across multiple concurrent productions
- Automated compositing workflows and render-farm optimization, improving throughput by 40%+
- Prototyped early ML integrations for asset tagging and pipeline intelligence

EARLY CAREER & ADDITIONAL EXPERIENCE

- **Pipeline Developer | Sony Pictures Imageworks** (2011–2012)—Integrated the studio's first Alembic geometry pipeline into the Hotel Transylvania production, working directly with Alembic's original authors
- **Stereoscopic Supervisor Employee #5 | StereoD** (2009–2011)—Pioneered foundational stereo-from-depth conversion workflows alongside the scientist who invented the technique as the company scaled from five people into one of the industry's leading stereo-conversion houses (acquired by Deluxe, 2011); the pipeline went on to power conversions for major studio franchises
- **Tech Lead—Automotive Visualization** (**Digital Giant**) · **Unity Developer—Spatial Localization SDK** (**Fantasmo**) · **VFX / Real-Time TD** (**Timber**, **Royale**, **House of Moves**, **Illuminate Hollywood**)

ADDITIONAL INFORMATION

Infrastructure — Operates production ML infrastructure including 4× NVIDIA B200 Blackwell GPUs and RTX A6000 for large-model deployment, fine-tuning, quantized inference, and reliability research.

Education — B.S., Computer Animation — Full Sail University