

Alpha Romer Coma *Machine Learning Engineer*

alpharomercoma@proton.me | alpharomer.com | linkedin.com/in/alpharomercoma | github.com/alpharomercoma

ML engineer building on-device LLM inference, PyTorch LLM training and post-training, and GPU and AI accelerator benchmarks.

TECHNICAL SKILLS

Training & Post-Training: PyTorch, Hugging Face Transformers, PEFT/LoRA, TRL, DPO, GRPO, torch.compile

Inference & Optimization: ExecuTorch, llama.cpp, vLLM, ONNX Runtime, MLX, GGUF and GPTQ quantization

Accelerators: CUDA (H100, H200), ROCm (MI300X), AWS Trainium/Inferentia, Google TPU, mobile NPUs

Software & Cloud: Python, Kotlin, TypeScript, SQL, Docker, Kubernetes, Terraform, AWS, Google Cloud

PROFESSIONAL EXPERIENCE

Experimental Machines , **Machine Learning Engineer**

Aug 2026 – Present

- Shipped **OpenWeights** to Google Play; its ExecuTorch LFM2.5 1.2B decodes **2.3× faster** than llama.cpp on Tensor G5.
- Published **16 ExecuTorch LLM exports** to Hugging Face for phone CPUs, GPUs, and Qualcomm and MediaTek NPUs.
- Built **ExecuServe**, an OpenAI- and Anthropic-compatible server that runs ExecuTorch LLMs from an Android phone.
- Merged a fix into **PyTorch ExecuTorch** that stops LFM2 models from carrying state into a new conversation.
- Fixed RoPE in MediaTek's stock NPU export for Qwen and Llama, lifting Qwen3 top-1 match with Hugging Face to **100%**.
- Secured a **\$2,000 compute grant** from Lambda for the team's GPU training and benchmarking work.

ViableView , **Senior Data Engineer**

Feb 2026 – Present

- Designed an email-enrichment pipeline processing **2.6M records at 240K/day** and achieving **0.94 macro-F1** on validation.
- Built ingestion for 2.7M registry records at 200K/day and an autoscaled multi-cloud pipeline validating 800K at 80K/day.

Kollab , **Associate Cloud Engineer**

Sep 2025 – Feb 2026

- Delivered a patient-search AI system for a Philippine hospital in **30 days** using Google Cloud, Vertex AI, and AlloyDB.

PROJECTS & RESEARCH

Small LLM Trained from Scratch to RLVR 

Aug 2026

- Pre-trained a **151M-parameter LLM** (GQA, RoPE, SwiGLU) on **8.26B tokens**, part of a 20-hour full pipeline on one H100.
- Implemented SFT, DPO, and GRPO in plain PyTorch; GSM8K answer format went from 5.5% after SFT to **98.5%** after GRPO.

LLM Fine-Tuning and Serving on AWS Trainium and Inferentia 

Aug 2026

- LoRA fine-tuned Llama 3.1 8B on Trainium at **68.3% MFU** (2,952 tok/s) and served the merged model with vLLM on Inferentia2.
- Matched CPU labels on all **10,000 ImageNet images** for compiled CLIP and SigLIP, and caught a CLIP graph returning NaN.

MI300X vs. H200 GPU Benchmark 

Jul 2026

- Measured 164 vLLM serving points; H200 ran 8B training steps **1.32–1.40× faster** despite MI300X's 1.32× FLOPs lead.
- Showed MI300X winning 70B FP8 serving at 128–256 concurrency by up to **1.14×** and fitting 70B in BF16 where H200 cannot.

Visual-Qwen: Multimodal Short-Form Video Classification 

Dec 2025

- Built a tri-modal EVA-CLIP/Whisper/Qwen3-4B VLM detecting “sludge” videos at **97.19% F1** on 300 held-out videos.
- Published the model, code, and a **2,000-video multimodal dataset** spanning TikTok and YouTube Shorts.

SPEAKING & COMMUNITY

International Conference Speaker, PyTorch Foundation & GitHub

Apr 2026 – Dec 2026

- Presenting “When the KV Cache Cannot Roll Back: Prefix Reuse on Android” at PyTorch Day Japan , Tokyo.
- Presenting “When Is Local Copilot Worth It?” at GitHub Universe , San Francisco.
- Presented “De-mystifying PyTorch for ASICs” at PyTorch Conference Europe , Paris.

PyTorch Ambassador , **GitHub Campus Expert** , **AWS Community Builder**

Dec 2024 – Present

- Selected as a 2026 PyTorch Foundation Ambassador to run events, create content, and mentor PyTorch users in the Philippines.
- Trained 4,000+ students and professionals in AI/ML, software engineering, and data engineering across 25+ talks.
- Secured \$30,000+ in sponsorships from GitHub, Tavily, and Vercel to support technology education across Luzon.

CERTIFICATIONS

- PyTorch Certified Associate (PTCA) 
- AWS Certified Machine Learning Engineer – Associate 
- Google Cloud Professional Machine Learning Engineer 
- Google Cloud Associate Cloud Engineer 

EDUCATION

BS in Computer Science with specialization in Software Engineering, FEU Institute of Technology 

Aug 2022 – Sep 2026

4-Year President's Scholar with High Honors; Leadership Award (3 of 921 graduates); CS Academic Varsity

AWARDS & RECOGNITION

AWS re:Invent 2025 All Builders Welcome Grantee , **AWS**: sole Filipino among 340 international recipients

Dec 2025

PH100 Honoree , **StellarPH**

Sep 2025

TESDA Academe Expert : validated two Philippine national AI and data competency qualifications

Aug 2025

National AI Prompt Design Challenge Champion , **Straits Interactive**

Jul 2025