

LUIS LACERDA

AI Systems Builder | AI Training Data & LLM Evaluation | Full-Stack Developer (TypeScript, Python)

New York, NY | luislacerda.site | linkedin.com/in/luis-lacerda-32a875405 | github.com/ABBYCRM

PROFESSIONAL SUMMARY

AI systems builder and software founder who designs, specifies, tests, and approves production code written with AI coding agents. Currently leading Mini-Mythos, an ongoing 36,000-step QLoRA fine-tuning run of Qwen3-14B (4-bit NF4, bf16 compute) on a single 12 GB GPU over a ~1.9 GB JSONL instruction corpus that includes frontier-coding-sft-v1, a curated 2,640-pair, 8-domain SFT component; the step-10,127 checkpoint (~28% of schedule) has been merged, converted to GGUF, and deployed for interim evaluation. Brings an LLM-evaluation mindset from daily work directing Claude Code, Grok CLI, Gemini CLI, and Cursor cloud agents under written specs, reviewing their output for correctness, and enforcing quality through 1,000+ API tests. Strong prompt engineering, technical writing, and domain expertise in LegalTech, mass tort and personal injury lead generation, and TCPA / CAN-SPAM / CASL compliance.

CORE SKILLS

- **AI Training & Evaluation:** SFT dataset creation (instruction/response pairs), QLoRA/LoRA fine-tuning, train/val splits, reviewing and approving AI-generated code, prompt engineering, multi-LLM orchestration, RAG (bge-small embeddings, Chroma)
- **Languages:** TypeScript, Python, C#/.NET, Rust, PowerShell
- **Frameworks & ML:** Node.js, Next.js, PyTorch, Ollama, LM Studio, ComfyUI/Flux, Kokoro TTS, GGUF conversion
- **Engineering Practice:** OpenAPI specs, API testing, GitHub PR workflow, DigitalOcean App Platform deploys, SEO and Lighthouse performance
- **AI Tools & APIs:** Claude Code, Grok CLI, Gemini CLI, Cursor cloud agents, MCP, Composio, Firecrawl, Exa, Tavily, xAI / Gemini / NVIDIA APIs
- **Business & Integrations:** HubSpot, Zapier, DocuSign, TrustedForm, Litify, Resend, Microsoft Teams, Google Workspace; Business, Data & Financial Analysis

EXPERIENCE

Founder | MTOS / ABBYCRM and TFT Legal Services

2026 - Present

Legal lead generation and LegalTech software for MVA, SSDI, VA disability, and mass tort practices.

- Direct a team of AI coding agents (Claude Code, Grok CLI, Gemini CLI, Cursor cloud agents): write the specifications, test the output, and approve changes through a GitHub pull-request workflow.
- MTOS CRM (TypeScript/Node): lead-distribution CRM for mass-tort firms with per-buyer lead validators, auto-send gating, duplicate tagging, Litify buyer integration, TrustedForm consent capture, an OpenAPI spec, and 1,000+ API tests.
- MTOS Portal: vendor portal with tracked referral links and live lead status.
- CaseClosedFL: Next.js personal injury platform; built lead validators for SSDI, CaseClosedFL, and F4MR.
- tcpa-firewall: TCPA list-scrubbing tool supporting compliant outreach.
- Launched ~12 legal lead-generation sites, including caseclosedmva.com and caseclosedfl.com, with multi-step qualifying forms, SEO, and Lighthouse scores of ~99; deployed on DigitalOcean App Platform.
- Manage the lead vendor network and LinkedIn outreach to law firms; apply TCPA, CAN-SPAM, and CASL compliance requirements.

ACS Neuro Network Architect | LI MEDIA

Jan 2014 - Present

- Apply business analysis, data analysis, and financial analysis across online advertising and SaaS work.

AI TRAINING & MODEL WORK

Mini-Mythos | Domain-adapted coding & reasoning LLM (QLoRA SFT)

In progress (Oct 2026 – present)

- **Training regime:** Parameter-efficient supervised fine-tuning of Qwen3-14B (abliterated variant) via QLoRA: frozen base weights quantized to 4-bit NF4, low-rank adapters optimized with bf16 compute (Hugging Face transformers + PEFT) on a single consumer GPU (RTX 4070 Ti, 12 GB VRAM).
- **Schedule & status:** Ongoing run on a 36,000-step optimization schedule. At step 10,127 (~28% of schedule), merged the adapter into the base weights mid-run, converted the merged checkpoint to GGUF, and served it via Ollama for checkpoint-gated interim evaluation while optimization continues.
- **Training data:** ~1.9 GB JSONL instruction/response corpus with a held-out train/validation split. Includes frontier-coding-sft-v1, a curated 2,640-pair component built in-house spanning 8 domains (incl. advanced coding, software engineering, cybersecurity).
- **Retrieval augmentation:** RAG over the same corpus: bge-small-en-v1.5 dense embeddings indexed in Chroma with top-k retrieval (k = 6), exposed through an OpenAI-compatible proxy with tool-use support.
- **Supporting work:** Directed creation of a PyTorch training guide (MNIST through LoRA/QLoRA and GGUF conversion); set up a Flux LoRA training method in ComfyUI; operate local inference stacks (Ollama, LM Studio, ComfyUI/Flux, Kokoro TTS).

AI SYSTEMS & AGENT ORCHESTRATION

- Designed and operate a fleet of 10+ coordinated AI agents (coding, code audit, recruiting, LinkedIn, Teams monitoring, Google indexing, media) with scheduled routines and watchdogs.
- Grok-Memory: git-based shared memory / RAG repository used across agents; GENERAL_RAG retrieval project.
- BOS-Omega: multi-LLM orchestration architecture. AURA-Runtime: agent runtime written in Rust.

ADDITIONAL PROJECTS

- LegalHeatmapApp (C#/.NET 8).
- Arabic speech recognition backend (quran-fastconformer).
- Robotics arm repositories (SO-ARM100, lerobot).