Karim Ramses Luna Ramos

Culiacán, Sinaloa, México
667 498 6938 | karim.luna.ramos@gmail.com | linkedin.com/in/karim-luna-ramos | github.com/karimluna

Profile

Chemical engineer and self-directed researcher working at the intersection of dynamical systems, optimization, and intelligent decision-making. Interests span the approximation of dynamical systems (SSMs, model order reduction), multi-agent coordination and control, and hardware-aware inference for resource-constrained deployment.

Education

M.Sc. in Automatic Control (candidate) Aug. 2026 – Present
CINVESTAV Guadalajara *Guadalajara, Jalisco*

- Research direction: approximation of dynamical systems through learned state-space models, with applications to real-time control

B.S. in Chemical Engineering – GPA: 8.4 Aug. 2021 – Apr. 2026
Universidad Autónoma de Sinaloa *Culiacán, Sinaloa*

- Thesis: benchmarked SAC, SMPC, DMPC, and Dynamic Programming for lithium-ion battery thermal management under realistic EV drive cycles; matched near-optimal DP performance at 100× lower online computation
- Coordinated a team of 6 for a year-long R&D program: *Artificial Intelligence in Chemical Engineering*

Experience

AI Engineering Consultant Apr. 2026 – May 2026
Confidential Client *Hermosillo, Sonora (Remote)*

- Unified three separate detector models into a single discriminative YOLO-based pipeline, improving average recall by 11.3% and reducing Docker image size by 1.5 GB
- Engineered model bundles preserving PyTorch and ONNX weights for continual learning from live plant data; integrated MLflow for centralized model registry and versioning
- Resolved critical tensor normalization bugs and implemented stratified sampling, stabilizing training on small industrial batches

Research Intern – Optimal Control and Process Optimization Jan. 2026 – Jun. 2026
TECNM / Instituto Tecnológico de Aguascalientes *Aguascalientes (Remote)*

- Studied and implemented Model Predictive Control (MPC) for constrained process systems under the supervision of Dr. Pablo Tenoch Rodríguez González
- Formulated optimal control problems using CasADi; compared deterministic MPC and stochastic MPC (SMPC) variants on uncertainty quantification benchmarks, GitHub.
- Applied receding-horizon control to battery thermal management, contributing directly to the RL vs. MPC benchmarking in the B.S. thesis

Data Scientist (Internship) May 2025 – Aug. 2025
Universidad Autónoma de Sinaloa *Culiacán, Sinaloa*

- Built high-frequency ETL system for industrial process monitoring; developed Gaussian and ensemble anomaly detection models reducing material waste by 13%
- Engineered feature pipelines improving model accuracy by 15% over rolling windows; built Grafana dashboards for real-time monitoring of critical-to-quality variables and model health

Process & Maintenance Engineer (Internship) Jun. 2024 – Aug. 2024
MASIPRO S.A. de C.V. – Toyota Motor Manufacturing Supplier *Querétaro, México*

- Led CAPA investigations for EHS process deviations using 5-Why and Ishikawa methodologies
- Applied data-driven analysis to identify system failures and support implementation of predictive maintenance from historical process data

Selected Projects

Mini-Mamba: Selective SSM from Scratch | *CUDA, C++, PyTorch, BF16* 2026

- Implementing the Mamba selective SSM (S6) from scratch: ZOH discretization, depthwise convolution, and a fused parallel selective scan kernel grounded in the associative scan pattern (GPU Gems 3, Kirk & Hwu)
- Training a 25M-parameter domain-specific language model in BF16 on an 8 GB consumer GPU; O(L) memory scaling vs. $O(L^2)$ for equivalent Transformer – Mamba runs at L=8192 where Transformer OOMs
- Fusing ZOH and state update into a single CUDA kernel, eliminating repeated HBM round-trips on a memory-bound operation (roofline AI $\approx$ 3 FLOP/byte)

Double Inverted Pendulum: RL Control & Embedded Policy Compression | *MuJoCo, PyTorch* 2026

- Trained a policy solving swing-up and balance of a double inverted pendulum (underactuated, 3-DOF, 1 actuator) in MuJoCo
- Benchmarked against LQR linearized at the upright equilibrium to demonstrate necessity of global RL. The policy recovers from arbitrary initial conditions where LQR fails outside its basin of attraction
- Compressed full policy (135K params) to embedded-deployable device (8K params, INT8) via structured pruning, knowledge distillation and post-training quantization; fits STM32H7 budget (32 KB SRAM, 50 Hz loop) while retaining balance stability under 2 N impulse perturbations

MADUS – Multi-Modal Multi-Agent Document Understanding | GitHub | *LangGraph, Docker, PyTorch* 2026

- Designed hierarchical orchestration over five specialized agents (text, tables, vision, critic, summarizer) with ReAct loop and Reflexion self-correction; hybrid BM25 + dense retrieval via Reciprocal Rank Fusion
- Applied post-training quantization (GPTQ, AWQ, SmoothQuant) to vision-language and text models; profiled memory bandwidth with nvml and torch.profiler, enabling 7B-class model inference within 8 GB VRAM budget
- Runs fully local via Ollama or scales to OpenAI backends; deployed with Docker, Redis, and n8n workflow automation

Tiny-GRPO: RL Load Balancer | GitHub | *Go* 2026

- Implemented Group Relative Policy Optimization as a zero-dependency, thread-safe load balancer in ~300 lines of pure Go; linear-softmax policy with 16 parameters and analytic gradients
- Converges from uniform routing to 87.8% on the healthy backend within 500 steps under simulated server degradation; same mathematical structure as MoE token routing and multi-robot task allocation

Technical Foundations

Mathematics: Real analysis, matrix analysis, convex optimization, control theory, linear systems, Bayesian filtering, approximation theory
Current study: Functional analysis, measure-theoretic probability, approximation of dynamical systems, CUDA kernel programming
Programming: Python, C++, Go, SQL, TypeScript (basic)
ML & Inference: PyTorch, JAX, TensorRT, ONNX, vLLM, Quantization (GPTQ, AWQ, SmoothQuant), LoRA/PEFT
Infrastructure: Docker, Kubernetes, FastAPI, gRPC, MLflow, Prometheus, Grafana, PostgreSQL, AWS (basic), Linux
Languages: Spanish (Native) | English (C1) | Mandarin Chinese (B1)