

LARBI SAID CHIKH

AI Engineer | LLM Agents, Verifiers & Evaluation

Valbonne, France | +213 553 988 691 | larbisk@gmail.com | github.com/larbi1512
linkedin.com/in/larbi-saidchikh-0621b0224 | researchgate.net/profile/Larbi-Said-Chikh

SUMMARY

AI Engineer building LLM agents for multi-step engineering workflows, and the verifiers and benchmarks that judge them by what actually happened on the system. At EURECOM, built an agent that writes and pre-checks 5G Core configurations, deploys them on a live cluster and repairs runtime faults (88.9% intent-to-deployment over 180 runs; CNSM 2026): the same generate-check-execute-repair loop that simulation and chip-design flows run on. Hands-on with SFT data and LoRA fine-tuning (NeurIPS 2025 workshop), PPO with custom rewards, and Python on Linux and GPUs.

EXPERIENCE

04/2026 – present	AI Engineer (part-time) Refactored the scaffolding of a production LLM data-science agent, then ablated prompt-engineering and prompt-compression variants to cut its context while preserving output quality. Built a reproducible analysis pipeline (custom parser, fully executed notebooks) over the PH-LLM fitness case-study data.	PrevAI Health
11/2025 – 06/2026	Research Intern (MSc Thesis) Designed an 11-agent LLM harness (LangGraph) that turns natural-language intents into Helm configurations, validates them (policy rules + Kubernetes dry-run), deploys a 5G Core (OpenAirInterface), then closes the loop: Prometheus KPIs → anomaly and SLA checks → LLM remediation plan → scale, restart, rollback or config patch. Covered by 225 unit and integration tests. Graded runs on deterministic ground truth instead of how outputs read: built a 36-intent benchmark over a complexity × load × scenario grid, scored resource plans against a formula-based oracle (no hand labels), and compared against 4 baselines, including a single-LLM agent. 88.9% intent-to-deployment over 180 live runs (36 intents × 5 repeats), zero human interventions. Failures were systematic, not random: 32 intents passed all 5 repeats and 4 failed all 5, three of them “simple” intents. The multi-agent pipeline cut config errors from 86.9% (single-LLM baseline) to 0% and remediated 30/30 injected faults (CPU spike, pod crash, traffic surge), vs. 0/15 with the remediation planner ablated. Benchmarked 8 open-weight LLMs (3B–120B, incl. Mistral-Small 3.1) as the harness backbone over 1,440 runs with paired, Bonferroni-corrected tests: accuracy jumps between 7B and 14B (0.38 → 0.69), then plateaus.	EURECOM, France / Remote
11/2025 – 03/2026	AI Development Intern Curated and validated an Arabic Q&A benchmark to evaluate a RAG pipeline whose keyword-filtering stage reduced retrieval noise (paper in preparation); shipped an OCR + LLM invoice-extraction service.	QANTRA
07/2025 – 09/2025	AI Research Intern Benchmarked BKT, DKT and Transformer knowledge-tracing models, lifting recommendation accuracy from 78% to 96%; ran attention and sequence-modeling ablations.	Siine
09/2024 – 11/2024	Geo AI Research Intern Deep learning on satellite imagery for land-cover classification and change detection.	Algerian Space Agency

PROJECTS

2024	RL for LLM Efficiency Cast three LLM-efficiency problems as RL tasks (prompt-token pruning, attention-head pruning, per-layer pruning + quantization) with multi-objective rewards trading quality (perplexity, output similarity) for latency, memory and FLOPs. On TinyLlama-1.1B, PPO reached 2.09% sparsity (1.100B → 1.077B non-zero weights).	github.com/larbi1512/RL-for-LLMs-optimization
2025	HPC Coursework: CUDA, OpenMP, MPI CUDA kernels (stencil, dot product, parallel scan) with 101–141× speedups over CPU, each verified against CPU results.	github.com/larbi1512/hpc-labs

PUBLICATIONS

2026	An Agentic AI Framework for Cloud-Native 5G Core Orchestration Multi-agent LLM framework for autonomous deployment and self-healing of a 5G Core on Kubernetes.	CNSM 2026 / Accepted
2025	Crafting Culturally Aligned Narratives: LLMs for Arabic Children's Stories H. Boukhalfa, S. Mani, L. Said Chikh. 714-story annotated SFT dataset (κ = 0.79–0.84), LoRA fine-tuning of four open LLMs (7B–13B) and Gemini 2.0, rule-based metrics plus a panel of 5 educators (ICC = 0.85).	MusIML @ NeurIPS 2025

EDUCATION

2021 – 2026	MSc, Artificial Intelligence and Data Science Thesis with EURECOM: <i>Agentic AI for Orchestration and Management of Telco Cloud</i>. Defended June 2026.	ENSIA, Algiers, Algeria
-------------	--	-------------------------

SKILLS

Code & ML: Python (primary, pytest), C/C++, SQL, Java; PyTorch, Hugging Face Transformers, LoRA, RL (PPO, Gymnasium)
Agents & evaluation: LangGraph, multi-agent orchestration, tool use, RAG, benchmark and reward design, LLM-as-judge, ablations
Systems & HPC: Linux, Kubernetes, Helm, Docker, Prometheus, Grafana, Redis, GCP, vLLM, Ollama, CUDA, OpenMP, MPI
Languages: English (C1, IELTS), French (C2, TCF), Arabic and Tamazight (native)
Community: mentor and speaker, GDSC AiFest'24 datathon; organizing committee, AI for Africa'23 conference