

JOSEPH AYANDA

AI/ML ENGINEER | CODING MODELS, EVALUATION & AGENTIC SYSTEMS

Ilorin, Nigeria | ayandajoseph390@gmail.com | josephmayo.site

github.com/josepha-mayo | huggingface.co/josephmayo

Independent AI/ML engineer building coding-model adaptations, verifier-driven post-training tools and practical AI applications. Work spans data curation, LoRA/QLoRA training, model merging, GGUF delivery, adversarial evaluation and multi-agent systems.

23 public model repositories

21,033 all-time model downloads

3 published datasets

Hugging Face public counters, 17 September 2026. Includes adapters and quantized releases; not unique users. Full breakdown on page 2.

SELECTED MODEL & RESEARCH WORK

Holo-3.1-9B-Coder | Coding-model adaptation

Fine-tuned Hcompany's Holo backbone and published the adapter, merged weights and Q4_K_M/Q5_K_M/Q6_K GGUF builds. Published evaluation reports HumanEval+ pass@1 of **65.2%** versus 52.4% base, and LiveCodeBench v2 of **37.8%** versus 31.5% base.

Gemma 4 E4B IT Coder | Training, evaluation and merge

Published a coding-tuned Gemma release with LoRA training and merge records from two Tesla T4 GPUs. Reported executable 50-task HumanEval-subset results improved from **34/50 to 42/50**; distributed merged weights, adapter and GGUF builds.

Holo 4B and LFM2.5 Coder | Compact and MoE adaptations

Released coding adaptations for a compact Holo model and LiquidAI's MoE model. Holo's published 80-task probe improved from **24 to 31 passes**; LFM's reported held-out mean negative log-likelihood decreased from **3.390 to 1.875**.

DR-OPIC | Verifier-driven post-training framework

Built a runnable Python framework that turns student coding failures into verified repair, preference and changed-span training records. Implemented curriculum weighting, rollout selection, deterministic replay, dataset auditing and optional PyTorch loss helpers for SFT, DPO and RLVR.

Model Unfetter | Directional-ablation research tooling

Implemented refusal-vector extraction, directional and norm-preserving ablation, per-layer/component controls and post-ablation evaluation. Added CPU/GPU, streaming and GGUF-oriented research backends.

TECHNICAL SKILLS

Models & data: PyTorch, Transformers, Hugging Face, PEFT, LoRA/QLoRA, SFT, DPO, ORPO, distillation, RLVR, GGUF, executable evaluation, synthetic data, filtering and deduplication.

Engineering: Python, TypeScript/JavaScript, C, Kotlin, FastAPI, Next.js, Node.js, PostgreSQL, SQLite, Redis, OpenCV, Docker, AWS Lambda, GitHub Actions and REST APIs.

EDUCATION

AI-Hikmah University, Ilorin | Cybersecurity undergraduate, in progress.

Model performance figures above are reported in the linked release cards. Subset evaluations and negative log-likelihood are not full-benchmark or independent replication claims.

OPEN MODELS & DATASETS

Release catalogue / cumulative downloads as of 17 September 2026

All model links below resolve to public repositories under **josephmayo**. Each format has its own count; family variants are grouped to distinguish full models, adapters and quantizations.

MODEL FAMILY / CONTRIBUTION	ARTIFACT DOWNLOAD COUNTS
Holo-3.1-9B-Coder Coding fine-tune; adapter, merged weights and GGUF in one repository.	All artifacts 4,237
ZAYA1-8B-Coder Custom-architecture coding adaptation and local inference releases.	Merged 132 GGUF 4,278 LoRA 63
Gemma 4 E4B IT Coder Coding SFT, full-weight merge and quantized distribution.	Merged 388 GGUF 1,752 LoRA 253
Holo-3.1-4B-Coder Compact coding model with an 80-task reported evaluation probe.	Merged 117 GGUF 813 LoRA 65
LFM2.5-8B-A1B-Coder Mixture-of-experts adaptation for multilingual coding tasks.	Merged 595 GGUF 1,330 LoRA 106
Qwopus-9B-Unfettered Directional-ablation research release and local GGUF builds.	Model 534 GGUF 2,815
Fara-7B-Abliterated-v2 Refusal-direction orthogonalization with a quantized companion.	Model 218 GGUF 600
Mellum2-12B-A2.5B-Thinking-Abliterated MoE reasoning-model research with per-expert and per-layer ablation.	Model 261 GGUF 705
Qwen2.5-0.5B-Unfettered Small-model ablation research for resource-constrained inference.	Model 1,440
HRM-Text-1B-sft-code Code-generation SFT experiment with a companion LoRA adapter.	Model 231 LoRA 73
Qwen2.5-agentic-7B-SLM-LoRA Adapter release targeting agentic and tool-use workflows.	LoRA 27

TOTAL: **21,033 model downloads** across 23 accessible repositories and 11 model families.

PUBLISHED DATASETS

Curated OpenBMB Code/Math | 197 all-time downloads

31,909 records: 25,891 direct-answer and 6,018 reasoning examples for code/math post-training.

Public Curated Coding Data | 226 all-time downloads

2,703 prompt/response rows across GitHub code, technical discussions and coding examples; source-specific licensing retained.

Refusal Compliance Pairs | 352 all-time downloads

201 labelled prompts: 101 refusal and 100 compliance prompts for activation contrast and directional-ablation research.

Download figures are dated Hub counters, not endorsements, paying users or de-duplicated installations. Model and dataset counts use different Hub counting rules. Research-ablation releases are listed as research artifacts, not safety-certified deployments.

SELECTED ENGINEERING PROJECTS

Open-source tools, application prototypes and reproducible experiments

ModelFang | Graph-based adversarial evaluation

Built a multi-turn LLM testing framework with conditional attack flows, finite-state response evaluation, provider adapters, execution budgets and audit logs; paired it with a Next.js analyst dashboard.

IntellectSafe | AI safety and security prototype

Developed a FastAPI/Next.js platform combining a multi-provider proxy, prompt and output review, privacy modules, multi-model validation and agent permission controls.

Mayo | Multi-agent repository maintenance

Implemented scanner, executor and reviewer agents that propose and review code changes, create GitHub pull requests and retain cross-repository lessons for subsequent runs.

SWARMs Debate Primitive | Multi-agent coordination

Built persona-based debate and voting workflows with transcript hashing and Solana-linked records, making the discussion and voting process inspectable.

Countback | OpenCV 5 and AWS inspection workflow

Built photo enrollment, viewpoint-aware review, evidence-linked notes and export. Evaluated 18 photos from nine fresh clips; validated the original analysis handler on private AWS Lambda with four valid analyses and five expected input rejections.

Counterstep and Pocket | Algebra repair and native Android practice

Developed typed repair and transfer-practice workflows that retain first attempts, hints and unfinished drafts. Shipped a Kotlin Android prototype with 98 JVM and 19 emulator UI tests in its recorded build.

SunQueue | Solar-aware work scheduling

Built a local JavaScript planner, dated shift sheets and a reported-run ledger. Enforced equivalent completed work and terminal battery accounting before displaying modeled grid-energy savings.

Forkline | Rollback and delivery-recovery lab

Integrated blockchain reorganization checks with a SQLite outbox and HTTP receiver. Demonstrated guarded dispatch, lost-acknowledgement reconciliation and restart recovery using an isolated local EVM.

ReturnReady | Merged open-source contribution

Built conservative comparison of written return terms and finished call results, with source-linked uncertainty and human review. Contribution PR #408 was merged into CALL-E's upstream repository.

DOOMFLY | Experimental vision and connectome interfaces

Built a physical fly-tracking setup and documented Doom/control experiments alongside a reproducible connectome-driven chess replay. Preserved protocol hashes, game records and measurement failures; did not label the chess run a live-fly result.

ADDITIONAL WORK

CutProof | Local-first clip review and source-context workflows. More code, model cards, demonstrations and technical notes are linked from [the portfolio](#).