

Aaweg Bhaladhare

Applied AI Engineer · Agent Benchmarks & RL Environments · LLM Evaluation · Agentic Systems

aawegwork@gmail.com · +91 92845 37320 · Pune, India (open to Bengaluru / remote) · [linkedin.com/in/aaweg](https://www.linkedin.com/in/aaweg) · github.com/Aawegg

SUMMARY

AI engineer who has built **400+ agent benchmark tasks** across terminal, multi-agent and enterprise-tool environments, plus the pipelines, QC gates and cloud runners that produce and grade them. Pod Lead for **LLM evaluation** at Turing (Terminal Bench, OS Bench, SWE-Bench). **Provisional Kaggle silver** (~top 1% of 10,246 teams) in an RL simulation competition, with a C++ simulator and PPO-trained policy.

EXPERIENCE

Turing

Jan 2025 – Present

Pod Lead, LLM Evaluation (Research Analyst until Sep 2025) · contract

Remote

Enterprise-tool agent benchmarks · Aug 2026 – Present

- Produced **300+ Harbor benchmark tasks** in which agents operate sandboxed **Slack, GitHub, Linear, Notion, Teams and Outlook** servers over **MCP**. Each is graded by deterministic **database-state and file verifiers** plus 1–2 rubric checks, with an oracle run required to hit **reward 1.0** and decoy answer paths proven to fail.
- Designed and ran the **LLM-agent mining pipelines** (Claude Code / opencode) behind them: seed selection from real client trajectories, substrate profiling, decoy design, verifier writing, per-figure **reachability proofs** from captured gym responses, and a final audit gate. Top committer to the team task registry (**706 commits**, 265 task records).
- One overnight batch: byte-exact evidence spans went from **6/20 to 20/20** and **36 blocking defects** were caught before shipping; a later batch passed **10/10** on all four gates (structure, oracle 1.0, reachability, client audit).
- Re-mine wave of **72 tasks**: all at oracle 1.000; **288 model trials**, 102 full passes (35%). Showed the dominant failure was agents paraphrasing required labels, so the score measured label compliance as much as difficulty, and reported it that way.
- Root-caused **0/4 model scores** to agent **context auto-compaction** at 131k tokens dropping the label list: label failures **61–75% with compaction vs 0% without**; tasks with ≤ 7 labelled lines passed 16/24 vs 2/40 for ≥ 10 . Disabled it for later waves.
- Traced a failure across every bundle to a **2-CPU container cap** against a hard-coded 40-second proxy boot budget while 12 services start. A/B on the same image: **dead at 46s capped vs ready in 18s uncapped**.
- Found the oracle replay driver could not represent an intentionally refused tool call: a task scored **0.12** with "required file missing" when the real cause was an access wall at step 6 of 21. Fixed the trajectory (**1.0, 9/9 verifiers**) and proposed an upstream **expect_error** flag.
- Built the **GCP Harbor runner fleet** (~40 scripts, 4.1k LOC): queue worker, enqueue / cancel / trial-kill, parallel-SSH multi-node dashboard with a web UI, an **SSE stream-filter proxy** that fixed orphaned gateway events crashing every agent run, and a patch for Harbor's hard-coded thinking flag.
- Also delivered **~95 synthetic tasks** across four batches and **56 seeded / seedless bundles**, triaging every model failure as MODEL, GRADER, INFRA or PROMPT before shipping.

Multi-agent benchmarks (SwarmBench) · May 2026 – Jul 2026

- Authored **18 multi-agent Harbor tasks** in fan-out, map-reduce and hierarchical patterns across competition math (HMMT, AIME, Putnam), regulatory research (CFR, IETF RFCs, World Bank) and data analysis; acceptance required the multi-agent run to beat single-agent by ≥ 0.23 .
- One task scored **0.32 single-agent vs 0.70 multi-agent**, rated "one of the strongest task designs in the batch" by the client.
- Built and open-sourced a **Codespaces harness** that runs oracle, single-agent and multi-agent trials on free GitHub Codespaces with Docker-in-Docker.

Pod Lead, LLM evaluation · Jan 2025 – Apr 2026

- Led a contributor pod on **Terminal Bench, OS Bench and SWE-Bench**: task distribution, review cycles, escalations and spec interpretation; also worked on OSWorld desktop-task QA.
- Extended a shared **QA reviewer for agent trajectories** (Google Apps Script, web + HTTP API): ~20 deterministic checks (degenerate loops, zero-tool-call runs, persona drift, PII and prompt leakage) feeding an **LLM-as-judge** on a weighted rubric, then a fixed PASS / CONDITIONAL / FAIL gate.
- Built a Python **LLM review pipeline** (OpenAI, Anthropic, Google Drive APIs) scoring task quality across **22 domains** with parallel batch reports, and tooling that converts solutions into **SFT notebooks** in assistant / tool-call / tool-output format.
- Directed rework of client-flagged tasks: average score **5.44 → 7.14** across 21 approved tasks. Authored a hard Terminal Bench task on **LoRA fine-tuning** of DistilBERT and a 545-line task-structure validator.

Independent AI Evaluation Contractor

May 2026 – Jul 2026

Terminal-Bench Task Author

Remote

- Authored **110 hard Terminal-Bench tasks** in Python and C++, each a **Dockerized environment** with a deterministic **pytest oracle** hardened against shortcut solutions; **61 accepted** by the client.
- Built the team's pre-submission tooling: a **local QC / LLM-judge gate** (~3.2k LOC) and a **FastAPI dedupe service** that fingerprints tasks, finds near-duplicates against past submissions (Jaccard over instruction and docs) and issues receipt tokens.
- Wrote a spec-authoring playbook for domain-original task specs after the client removed templated submissions, targeting hard tasks at $\leq 20\%$ model pass rate.

- Built a **YOLOv8** object-detection model for mine-worker safety monitoring, reaching **95% accuracy** on protective-gear detection.
- Shipped a **face-recognition attendance system** for **50+ employees** and contributed to the model-serving / inference-API layer of a full-stack LMS (Next.js, React, MySQL).

PROJECTS

Kaggriculture (Kaggle) — Provisional Silver, ~top 1% of 10,246 teams

Aug – Oct 2026

C++ · Python · PyTorch · PPO · behaviour cloning · cloud CPU / GPU runs

- Built a **C++ replica of the game engine** with move-for-move parity to the official one at **~1.5M games/hour**; every later experiment ran on it.
- Trained a late-game policy with **PPO** on final money difference: **+\$400/game**, held on 2,400 fresh games. Neural-policy ensemble members reached ratings of **2830 / 2831** (top-10 range at the time).
- Behaviour cloning on **850k decisions** from the #1 team hit **81% offline accuracy** but earned \$52 in a live game vs \$26,505 for my hand-written bot; a 2551-rated agent turned out to have won 20% of its last 30 games. Rebuilt evaluation around live opponents.

Benchmark Task Generation Pipeline

Sep 2026

Python (stdlib) · Docker · uv · ruff · pytest · Claude API with replay caching · mutation testing

- Repo-agnostic pipeline: audits a Python repo, builds a **hash-pinned lockfile** through a resolution ladder, a **digest-pinned multi-arch Docker** image and lint config, then mines tests and benchmark tasks with an LLM.
- On glom: **90 generated tests** (suite 266 → 356) killing **70 of 72 injected mutants**; **10 validated tasks** that fail on the parent commit and pass on the fix; two clean builds gave byte-identical per-test results. Rehearsed on a second repo (humanize).

NibbleCore — 4-bit Quantization Kernels for Apple Silicon

2026

C++ · ARM NEON SIMD · GGUF · llama.cpp · [github.com/Aawegg/nibblecore](#)

- From-scratch **Q4_0** quantize / dequantize kernels and a **GGUF parser**, benchmarked on real SmolLM2 weights: **3.4x faster dequantization** than llama.cpp's scalar reference and a **4.8x faster Q4xQ8 SDOT** dot product, bit-identical output; 7.1x compression at MSE 0.0016.

priorityjudge — LLM-as-Judge Plan-Quality Auditor

2026

Python · Gemini · Pydantic · scipy · GitHub Actions CI · Docker · [github.com/Aawegg/priorityjudge](#)

- Multi-pass agent (extractor → 5 per-criterion scorers → **deterministic citation verifier** → synthesizer). On a 6-plan, 3-repeat calibration set: **9887 / 10000, perfect Spearman** vs the human grader, **96.5% citation precision, 99.6% test-retest stability**; ~30 CI tests.

MosaicMind — Multimodal RAG & MLOps Pipeline

2026

FastAPI · LangGraph · LlamaIndex · ChromaDB · Gemini · CLIP · Groq Whisper · MLflow · Airflow · Docker · [GitHub](#) · [live demo](#)

- Agentic **multimodal RAG** over PDF, image, audio and video: a **LangGraph** plan → retrieve → synthesize → reflect loop over two ChromaDB stores (Gemini text + CLIP image embeddings) with SSE streaming.
- **MLflow** tracking per query, **Airflow** re-index and eval DAGs, pytest + ruff CI; live on HF Spaces, Cloud Run deploy scripts.

OPEN SOURCE

Rthesis (AI testing platform) — rejected execution of empty test sets before Celery scheduling, with tests

PR #2173 · merged

OGX / Llama Stack — fixed Gemini provider auth docs (mutually exclusive api_key / access_token)

PR #6341 · merged

SKILLS

Languages	Python · C / C++ · SQL · Bash · TypeScript · Google Apps Script
Evaluation & RL	Harbor · Terminal Bench · SWE-Bench · OS Bench · SwarmBench · RL-environment design · deterministic verifiers · LLM-as-judge · rubric design · PPO · behaviour cloning · mutation testing
LLMs & Agents	MCP · tool calling · multi-agent orchestration · LangGraph · LangChain · LlamaIndex · RAG · Claude Code / opencode agent pipelines · context engineering
ML / DL	PyTorch · Hugging Face Transformers · LoRA · CLIP · YOLOv8 · OpenCV · quantization (GGUF, ARM NEON)
Infra & MLOps	Docker · GCP (Compute Engine, GCS, Artifact Registry, Cloud Run) · GitHub Actions · MLflow · Airflow · FastAPI · Pydantic · pytest · Linux

EDUCATION & CERTIFICATIONS

B.Tech, Artificial Intelligence & Data Science — Vishwakarma Institute of Information Technology, Pune

2021 – 2025 · CGPA 8.54

Certifications: IBM Data Science Specialization · IBM Machine Learning · Microsoft Power BI Data Analyst · Databases and SQL for Data Science (IBM) · Career Essentials in Generative AI (Microsoft & LinkedIn)