

Yasir RAZA

✉ yasirabdali6@gmail.com | 🌐 github.com/Yasirrazaa | 🔗 linkedin.com/in/yasir-raza | 📍 Islamabad
☎ +92 314 5363379 | 🐍 pypi.org/user/yasirabdali6 | 🏆 kaggle.com/yasirabdaali | 🌐 yasirrazaa.github.io



SUMMARY: Research Engineer focused on LLM training and evaluation methodology: rule-based RL (**GRPO/RLVR**), execution-based reward design, contamination-checked benchmarks, and adversarial robustness. Built an oracle-verified RL training harness; my Archon security platform ranks **13th/100+ overall** (49.7% avg; 85.1% defense / 14.4% attack) in Berkeley's AgentBeats Arena.

TECHNICAL SKILLS

LLM TRAINING: GRPO / RLVR (TRL), supervised fine-tuning, reward design & shaping, curriculum construction, LoRA/QLoRA (PEFT, Unsloth), PyTorch

EVALUATION METHODOLOGY: Execution-based harnesses, SHA-256 frozen splits, bootstrap 95% CIs, pass@k, benchmark design (AgentDojo, GAIA)

ADVERSARIAL ROBUSTNESS: Prompt-injection attack/defense, red-teaming, GOAT/PAIR/Crescendo taxonomies, defense gateways, OWASP ASI01–10

SYSTEMS & INFRA: Sandboxed code execution (AST allowlists, RLIMIT), CUDA graph optimization, Docker, CI/CD, FastAPI

LANGUAGES: Python, SQL, Bash, TypeScript

WORK EXPERIENCE

TEAMLIFT — UNITED STATES (REMOTE).....

Junior Data Scientist

Jan 2024 - Dec 2024

- Scraped **thousands of AI tools** from public tool directories; built an automated verification pipeline (status codes, screenshots, freshness checks) that kept the dataset powering Teamlift's **Teamo** recommendation engine trustworthy.
- Preprocessed and normalized thousands of tool records into model-ready features, eliminating stale and duplicate listings.

Data Science Intern

Jul 2023 - Dec 2023

- Built an automated pipeline collecting **thousands of job postings** from dice.com with deduplication; extracted skill-demand signals from postings.
- Trained **time-series models** to forecast skill demand, treating label quality and measurement as first-class evaluation problems.

RESEARCH PROJECTS

Math2Code — Rule-Based RLVR for LaTeX → Python

2025 - 2026

- Built the complete verified training + evaluation stack for a math LLM: **22,796 deduplicated samples** with a SHA-256 frozen split, oracle-verified gold solutions (**397/397**), bootstrap-CI execution metrics, **no LLM-judged rewards** (GRPO rewards on freshly resampled inputs). Sandbox: **10,000 execs in 26s**.

Archon 1.0 — Adversarial Agent Security (Competition)

2026

- Scenario-agnostic attacker/defender agents (A2A + Google ADK): **13th/100+ overall (49.7%)**, **6th defense (85.1%)**, **18th attack (14.4%)** on the private leaderboard.

Archon 2.0 — Measurable Attack-and-Defense Platform

2026

- The only open platform where an adaptive attacker and a measurable defense fight in the same loop: **222 probes** across 10 packs, full **OWASP ASI01–10** coverage, AgentDojo v1 benchmark (0% ASR on direct overrides at the zero-cost tier), CVSS-style severity — enforced by **2,295 tests** (93% gate).

PRISM-Bench — Cultural Intelligence (CQ) Benchmark

2026

- Authored **PRISM: Pluralistic Reasoning & Identity-Specific Modeling**, an open benchmark measuring cultural intelligence in AI systems across pluralistic scenarios.

EDUCATION

B.Sc. Software Engineering, National University of Modern Languages (NUML), Islamabad

Aug 2026

CERTIFICATIONS

CERTIFICATES: Kaggle **2× Expert** (Datasets · Notebooks, 13 medals) · DataCamp DS/DA career tracks · 4 Google Cloud Gen AI badges · NUST LLM Bootcamp