Vyomakesh Dundigalla

AI Researcher | Agents | Evaluation Infrastructure | Inference Engineering | GPU Systems

vyomakesh018@gmail.com | linkedin.com/in/vyomakesh-dundigalla-061575159
github.com/vyomakesh0728

Summary

AI research engineer focused on agents, post-training alignment, reproducible evaluation systems, inference engineering, and GPU-level performance optimization. I build agent/eval infrastructure, verifier-backed training loops, trace instrumentation, and practical systems for turning research workflows into reliable experiments.

Research Interests

- Reinforcement Learning (RL), self-distillation, pre/post-training, reward modeling, and preference optimization.
- Evaluation frameworks for agents, benchmark design, and instrumentation for failure analysis.
- Inference engineering for LLM serving, rollout generation, and performance-aware deployment.

Highlighted Achievements

- Optimized Zerobrew package manager downloads 3.6x faster (6 min to 1.7 min) via racing cancellation and HTTP/2 tuning (Jan 2026).
- Achieved 12.804 microseconds in NVIDIA GPU Mode "nvfp4 block-scaled GEMM" optimization hackathon on B200, targeting speed-of-light performance with CuTeDSL (Dec 2025).
- Built 6 reproducible evaluation environments and benchmarks (2 safety, 2 web, 1 tool-use/MCP, 1 game suite) for PrimeIntellect verifiers (2025).

Experience

LiteFold **Apr 2026 – Present**
Founding AI Researcher *Bangalore, India*

- Building PLM training and inference pipelines for protein design, from sequence proposal optimization to model reasoning systems.
- Building verifier-backed reward infrastructure for PLM workflows, turning model outputs into computable training and evaluation signals.

Zerobrew **Jan 2026**
Performance / Systems Engineer (OSS) *Remote*

- Delivered 3.6x download optimization (361s to 99s) with racing cancellation and HTTP/2 tuning, eliminating redundant bandwidth usage through atomic coordination (AtomicBool, Notify, and Semaphore).
- Merged to main and immediately impacted all users.

PrimeIntellect-ai (Environments Hub) **2025 – Present**
AI Research Engineer (Agents & Evals, OSS) *Remote*

- Core contributor to evaluation infrastructure with 5+ merged PRs spanning benchmarks, web agents, and training pipeline integrations.
- Built 6 evaluation environments across web agents, tool-use systems, and games with a unified API contract for reproducible cross-benchmark testing.
- Separated rollout, verifier, and scoring paths across benchmarks to make failures easier to attribute during evaluation runs.
- Added sandboxed execution boundaries and process isolation for safer, auditable agent runs.
- Key contributions: BALROG game suite (6 simulators), BrowsecompPlus, WebVoyager, JailbreakBench and WMDP safety evals, and HuggingFace MCP integration.

Proposals

VGAC: Verifier-Grounded Agreement-Calibrated RL **2026**
Independent research proposal

- Proposed grouped rollouts, verifier signals, and entropy-aware calibration to reduce silent overconfident failures in long-horizon agents.
- Connected training-time reward shaping with inference-time verify/replan routing for agents that need robust uncertainty estimates.

Technical Skills

Languages: Python, C, C++, Rust, Bash, SQL, CUDA
DL / ML: PyTorch, TensorFlow, Keras, NumPy, Matplotlib, Pandas, SciPy
LLM Training & Serving: Transformers, PEFT, vLLM, FSDP, Ray, verl, DeepSpeed ZeRO
GPU: Cutlass, CuTeDSL, FP4/FP8 kernels, throughput and latency optimization
Eval & Agents: Stateful, web, and tool-use benchmarks, reward design, trace instrumentation
Infra & Tooling: Docker, sandboxed execution, API integrations, profiling and performance tooling
Web & Others: HTML, CSS, JavaScript, FastAPI, Node.js, Git, LaTeX

Education

Queen Mary University of London — **Sep 2022 – Oct 2023**
MSc in Business in Blockchain & Society — *London, UK*

Keshav Memorial Institute of Technology (Aff. JNTU) — **Oct 2017 – Sep 2021**
Bachelor of Technology in Electronics and Communication Engineering — *Hyderabad, India*

Certifications

Deep Learning Specialization | *DeepLearning.AI* — **Apr 2020**
Google IT Automation with Python | *Google* — **Aug 2020**