HRIJUL DEY

AI & Backend Engineer — Production Agents, APIs, and Measurement

Thane (Mumbai), India · Remote-friendly, open to hybrid | phant0mhriju@gmail.com | +91 62908 18033 |
https://www.linkedin.com/in/dey-h-ml972 | github.com/hr1juldey

SUMMARY

AI engineer since 2019, starting in computer vision and gait analysis on the CMU motion-capture database in biomedical R&D. In production LLM systems since February 2025, now owning "Kai", the agent platform a 500-person organization runs daily: hundreds of federated tool servers, routing, token budgets, failure isolation. Cut organization-wide LLM token cost ~93% — a 15x reduction — while reliability held as usage scaled to org-wide default. Trained small task-specific models (16–200 MB) for tool selection, grouping, and ordering — an experimental run, rolled back when measurement favored deterministic routing. My public evaluation harness and benchmark live at https://github.com/hr1juldey/ctx-gateway.

CORE SKILLS

- Backend: Python (production daily), APIs, async services (asyncio/anyio), Linux, Docker, Git, CI gates
- AI systems: MCP servers and gateways in production, tool routing, task decomposition, multi-server orchestration, evals and regression gates
- LLM engineering: PyTorch, training and fine-tuning small open-source models, quantization, prompt optimization, local inference (llama.cpp, Ollama)
- Data: PostgreSQL, Redis, data pipelines, statistical comparison of production distributions
- Foundations: computer vision (OpenCV, MediaPipe), gait analysis, measurement systems, regulated-domain validation

PROFESSIONAL EXPERIENCE

Product Engineer, AI Platform — Mosaic Wellness | Thane, Mumbai | Jun 2026 – Present

- Own "Kai" end-to-end: an MCP gateway federating hundreds of tool servers behind one governed endpoint, serving a 500-person organization daily.
- Define and track agent quality: task success rate, cost, reliability, stability — measured on production traffic, versioned in Git behind reproducible regression gates.
- Cut organization-wide LLM token consumption ~93% (a 15x reduction) while reliability held as usage scaled to org-wide default.
- Tune agent behavior via reward-guided steering at the harness layer — prompt optimization against production objectives — without modifying model weights.
- Trained and fine-tuned small open-source models (16–200 MB, tightly parameter-targeted for minimum compute) to automate tool selection, grouping, and ordering inside Kai — shipped experimentally, then rolled back deliberately when measurement favored deterministic routing.
- Built the failure-handling loop: production faults trigger an automated, agent-driven fix pipeline that scopes and ships micro-fixes, gated by adversarial tests before merge.

Product Manager & Software Engineer — Lilypad (EV marketplace) | India | Feb 2025 – Apr 2026

- Shipped production LLM features on India's first MCP-driven marketplace: tool integration, APIs, data flow.
- Owned requirements through release; prioritized by business impact and technical feasibility.

R&D Specialist, Computer Vision & Data Science — Ashva (biomedical devices) | Vellore | Dec 2019 – Jan 2023 (concurrent with final B.Tech years)

- Estimated joint loading (knees, ankles) from multi-sensor motion data, including immature, partially-calibrated sensors — multi-sensor correlation analysis on the CMU motion-capture database.
- Computer vision for posture detection and gait analysis.
- Test design, validation, and documentation against regulated standards.

Summer Intern — Variable Energy Cyclotron Centre (DAE) | Kolkata | Apr – May 2019

OPEN SOURCE (active, public — 22 PRs across 6 orgs in four days, Nov 2026)

- smolagents (Hugging Face): token-accounting fix MERGED upstream (PR #3206); parser PR open. https://github.com/huggingface/smolagents/pull/3206
- Trail-Code (trailhq, TypeScript): 13 PRs under review — search, guards, MCP tool matching; each tied to a filed issue, tests, CI green. https://github.com/trailhq/Trail-Code/pulls?q=is:pr+author:hr1juldey
- hermes-agent (NousResearch): 5 PRs open — credential-store redaction, host-file read boundaries for plugin backends, a terminal safety floor that holds under force overrides, model-preset caching, cron-recovery docs. Found by running the agent daily on itself.
- compos (svs): 2 PRs + 4 filed issues (daemon robustness, model discovery); tau (Hugging Face): forked, auto-compaction bug reported upstream.

ctx-gateway — https://github.com/hr1juldey/ctx-gateway

- Clean-room MCP gateway with reproducible benchmark gates and a public eval dataset under construction: tool-definition payload in context, task success rate, p95 latency. Runnable by anyone; build log records what broke and how it was fixed.

VerifAI — https://github.com/hr1juldey/VerifAI

- Fraud detection with self-supervised representation learning (JEPA-class); evaluation-first design with precision/recall gates.

INDEPENDENT PROJECTS (personal)

- Training a tumor-detection model for a medical-imaging client (2026, in progress).
- Physics-validated motion pipeline on the same CMU database I first used in 2019: acceptance criteria are measured physical invariants (foot-slide, ground penetration, COM acceleration).
- Local multimodal inference on a 12 GB consumer GPU: LLM + vision-language model co-resident under a shared VRAM budget; fine-tuning and quantization of small open-source models.

EDUCATION

B.Tech, Biotechnology — Vellore Institute of Technology (2018–2022)

CERTIFICATIONS (Google)

Introduction to Large Language Models | Introduction to Generative AI | Generative AI Explorer — Vertex AI

LANGUAGES

English (native/bilingual), Bengali (native), Hindi (native), Spanish (limited working), French (elementary)