Divyansh Agrawal

Applied AI Engineer | Production AI Systems | Agents & Evals | AI Infrastructure

+91 98191 44432 | divagr1925@gmail.com | linkedin.com/in/divyansh-agrawal-03476724b | github.com/divagr18 | divagr.com

Experience

BizDateUp Jul. 2026 – Present
AI/ML Technical Lead *Mumbai, India*

- **Built an agentic RAG system for startup due diligence** over ∼500 documents each (idea, SWOT, financials, legal); cut first-pass review from 1–2 weeks to ∼**4 hours** and **halved external CA spend** across **30 startups**.
- **Took an automated GTM system to production** for lead sourcing, investor matching, and outreach, integrating LinkedIn Sales Navigator and RocketReach to surface **300 prospective investors/day** ranked by a tuned rubric.
- **Manage 2 engineers and advise 5 portfolio engineering teams on AI/ML**, including architecture and evals.

Bajaj Finserv Health Dec. 2025 – Jun. 2026
Software Engineer Intern *Pune, India*

- **Maintained and upgraded an AI adjudication system for health insurance claims**, reporting to the CTO; processed **50k+ claims and 1M+ documents/day** at **97% accuracy**.
- **Owned and scaled the Kafka-driven claims platform** processing 1M+ claims/week on 60 workers across 12 VMs; cut **infra cost 64%** by moving to Azure VMSS, cut **LLM API cost 73%**, and cut **DB query latency 47%**.
- **Productionized a real-time STT–LLM–TTS agent stack** with tool use, PII redaction, prosody control, batching, KV-cache reuse, and Q8 inference across 10+ STT/TTS models at **76% lower serving cost**.

Achievements

Winner	**Hack Apertus Online:** red-teamed Apertus, the Swiss open LLM, and found behavioural faults
Winner	**Cerebras × Google DeepMind Gemma 4 Hackathon:** Trace, autonomous debugging agent
Winner	**ISB Hackfest:** on-prem voice clinical AI for documentation, ICD-10 coding, and insurance
Finalist / 30,000+	**Meta × PyTorch × Hugging Face:** GRPO-trained Kubernetes agent
Ambassador	**Qwen Developer Program, Alibaba Cloud:** demos, open-source work, and model feedback

Projects

Qwenputer | GitHub | *Compiled Transformer Weights, Model Internals, Tool Use*

- **Compiled a 32-bit RISC-V CPU into the weights of Qwen3.5-4B**: 11 added layers with computed weights run C and MicroPython inside the forward pass and pass the official riscv-tests suites.
- **Raised accuracy on DeepMind Mathematics arithmetic with no external tool**, such as big-number multiplication from 55.0% to 97.5% and big-number division from 72.5% to 97.5%.

FactorioRL | GitHub | *Reinforcement Learning, Simulation, Sim-to-Game Transfer*

- **Built a tick-exact Factorio simulator at 947k decisions/sec on 16 CPU threads** (∼3,600× the real engine) with trace-level parity, driving RL training at 26,800 steps/sec on an RTX 4060.
- **Transferred simulator-trained policies to the real game**: agents trained only in simulation reach **90.6% success** on unseen scenes in real Factorio, and held-out success on two build tasks rose from under 2% to **87–95%**.

trunc | GitHub | *Rust, Systems Programming, Crash-Safe File I/O*

- **Built a Rust extractor that unpacks archives bigger than the free disk space**, deleting the archive as it extracts; cut peak disk use **37%** vs. 7-Zip and Windows while running faster than both.
- **Designed a back-to-front archive format and crash tests at every filesystem operation**; it predicts peak disk use within 2% and resumes after crashes or full disks.

EchoNet | GitHub | Preprint | *Agent Evaluation, Epistemic Robustness, Retrieval*

- **Built a simulated internet and search engine to test whether AI models can be misled** by what they read online: planted false pages, many sites echoing the same false claim, and sources with misleading origins.
- **Scored models on whether they resist misinformation yet still change their answer** when genuine new evidence arrives, alongside how faithfully they cite sources and how well their confidence matches reality.

Technical Skills

AI/ML: LLM Agents, RAG, Evals, Post-Training & RL (PPO, GRPO, RLOO), Quantisation, Inference Serving, MCP
Frameworks: PyTorch, Transformers, vLLM, llama.cpp, FastAPI, Django, Node.js
Data & Infrastructure: PostgreSQL, pgvector, Redis, Kafka, Kubernetes, Docker, CI/CD, Cloudflare, AWS, GCP
Languages: Python, TypeScript, JavaScript, SQL, Bash

Education

Mukesh Patel School of Technology Management & Engineering Mumbai, India
B.Tech in Computer Engineering with Honors in AI/ML *2022 – 2026*