

Divyansh Agrawal

9819144432 | divagr1925@gmail.com | linkedin.com/in/divyansh-agrawal-03476724b/ | github.com/divagr18 | divagr.com

EXPERIENCE

Qwen / Alibaba Cloud

May 2026 – Present

Qwen Dev Ambassador

Remote

- **Represent Qwen across the AI developer ecosystem** through technical demos, open-source contributions, community content, and model advocacy.
- **Create developer-facing examples and technical content** around Qwen models, workflows, and deployment.

Bajaj Finserv Health

Dec. 2025 – Present

SDE Engineer Intern

Pune, India

- **Built and optimized a real-time streaming AI voice stack (STT–LLM–TTS)** with tone and prosody control, PII redaction, and custom toolchains; deployed 10+ STT/TTS and full-duplex models on H100 GPUs using batching, KV-cache reuse, and Q8/BF16 precision to support high-concurrency sessions at **76% lower cost**.
- **Owned and scaled a Kafka-driven platform processing 1M+ claims/week** on 60 workers across 12 VMs with ~20-stage, multi-variant workflows; implemented idempotency, back-pressure control, recovery, and observability.

ACHIEVEMENTS

- **Winner, Cerebras × Google DeepMind Gemma 4 Hackathon:** Built Trace, a multimodal support-to-engineering system that completed triage, reproduction, source localization, patching, verification, and GitHub handoff in ~12 seconds using Cerebras inference at up to 1500 TPS.
- **1st Place, ISB Hackfest:** Built an on-prem agentic system for voice-based clinical documentation, ICD-10 treatment coding, and insurance workflows; judged across prototype, technical build, product, business plan, and GTM.
- **Finalist / 30,000+ teams, Meta × PyTorch × Hugging Face Hackathon:** Built a GRPO-based reinforcement-learning setup to train Kubernetes-management agents under cost and latency constraints.
- **National Rank 2 / 15,000+ teams, Bajaj HackRx 6.0:** Built an optimized RAG-based agentic system with sub-second retrieval and reranking, plus an autonomous browser-use agent for web-grounded reasoning.

PROJECTS

Trace | [GitHub](#) | *TypeScript, Cerebras, Gemma 4, Playwright* | *Cerebras × Google DeepMind Winner*

- **Built Trace, a multimodal support-to-engineering agent** that turns vague bug reports into reproduced, patched, and verified GitHub PRs in ~12 seconds using Cerebras-hosted Gemma 4 31B.
- **Orchestrated the end-to-end debugging pipeline** across multimodal triage, Playwright mobile reproduction, source localization, constrained patching, regression checks, and automated GitHub handoff.

Memlayer | [GitHub](#) | *Python, LLM Memory, Retrieval (250+ stars, 3k+ weekly PyPI downloads)*

- **Created a lightweight memory-layer framework** for LLM applications with structured long-term memory, fast contextual retrieval, and pluggable storage backends, adopted across thousands of installs.
- **Built a plug-and-play Python SDK** for adding inference-time memory to existing agent and retrieval pipelines.

Hyperion | [GitHub](#) | hyperionhq.co | *Go, LLM Gateway*

- **Built a high-performance AI gateway** handling ~20k requests/sec with ~5μs routing latency across provider abstraction, model selection, budgets, and failover.
- **Implemented two-layer semantic caching, cost-aware routing, and air-gapped PII redaction** at ~35μs p99 with ~3μs failover.

EDUCATION

Mukesh Patel School of Technology Management & Engineering

Mumbai, India

B.Tech in Computer Engineering with Honors in AI/ML

2022 – 2026

TECHNICAL SKILLS

Languages & Frameworks: Python, Go, Lua, SQL, JavaScript, TypeScript, React, Django, FastAPI, Express
Backend & Systems: System Design, Kafka, Redis, Kubernetes, Docker, CI/CD, PostgreSQL, AWS, Azure, GCP
AI/ML Systems: LLM Inference, Reinforcement Learning, Post-Training, Quantization, CUDA/Triton Kernels, Attention/KV-Cache Optimization, Speculative Decoding, MoE Inference, Agentic & Multimodal Systems
Certifications: Regression, Classification and Clustering, Deep Learning, Statistical Inference, Computer Vision