Divyansh Agrawal

AI/ML Systems | LLM Research | Inference | Agents | Alignment | Evaluation

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

Selected Research

WebBench / EchoBench | GitHub | Paper | *Agent Evals, Epistemic Robustness, Retrieval*

- **Built a controlled synthetic-web benchmark for epistemic arbitration** with 600 matched counterfactual worlds testing poisoning, manufactured consensus, source provenance, and legitimate belief updates.
- **Designed hybrid retrieval, hidden provenance graphs, deterministic scoring, and calibration metrics** for reproducible cross-model evaluation of web-connected agents.

SpectralQuant | *Quantization, Model Compression, Activation-Aware Calibration*

- **Developed an activation-aware low-bit quantization framework** that shapes Q4 error around downstream activation damage rather than weight-space reconstruction alone.
- **Recovered 96.5% of the BF16-to-Q4 loss gap on Qwen3.5-0.8B** at the same 4.52 BPW footprint, with gains replicated across Qwen, Gemma, and Llama experiments.

StateShift-JEPA | *World Models, Representation Learning, Agentic Reasoning*

- **Developed an action-conditioned latent world model** that predicts future state representations to detect stale agent beliefs after hidden environment changes and trigger repair.
- **Built StateShift-Bench and a predictor-repair planner** with hidden-shift TextWorld environments, uncertainty-aware action scoring, and frontier-model baselines.

Model Spec Midtraining Order | GitHub | *Alignment, Post-Training, Model Behavior*

- **Study whether curriculum order drives alignment generalization** by adding the missing demonstrations→spec arm alongside spec→demos, demonstrations-only, spec-only, and shuffled controls.
- **Run preregistered LoRA experiments and OOD alignment evaluations** across Gemma and Llama using preference, open-ended, capability, and learned value-direction analyses.

Selected 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.

Selected Projects

Trace | GitHub | *TypeScript, Cerebras, Gemma 4, Playwright* | *Cerebras × Google DeepMind Winner*

- **Built a multimodal autonomous debugging agent** that turns vague bug reports into reproduced failures, source-localized patches, verified fixes, and GitHub PRs in ∼12 seconds using Gemma 4 31B on Cerebras.
- **Designed a dependency-aware agent orchestration loop** spanning multimodal triage, browser reproduction, code reasoning, constrained patching, regression checks, and automated GitHub handoff.

Pull Guard | GitHub | Demo | *Agent Evals, Code Intelligence, Adversarial Verification*

- **Built an AI-native maintainer control plane for large PR queues** combining semantic change clustering, source-grounded review, patch–claim alignment, and adversarial test generation.
- **Designed a self-verifying execution pipeline** that validates generated tests against clean, candidate, and deliberately broken revisions inside isolated disposable runners.

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

- **Built an inference-time memory and retrieval layer for LLM agents** with structured long-term memory, hybrid vector/graph retrieval, and pluggable storage, adopted across thousands of installs.
- **Designed a model-agnostic Python SDK** for injecting persistent context into existing agent and RAG pipelines.

Hyperion | GitHub | hyperionhq.co | *Go, LLM Gateway*

- **Built a high-performance LLM inference gateway** handling ∼21k requests/sec with ∼5μs routing overhead across multi-provider model routing, budgets, and failover.
- **Implemented two-layer semantic caching, cost-aware model routing, and air-gapped PII redaction** at ∼35μs p99 with ∼3μs failover.

Experience

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

- **Lead AI/ML development across investment and portfolio workflows**, designing agentic systems for research, diligence, document intelligence, and operational automation.
- **Build production LLM pipelines spanning retrieval, tool use, structured reasoning, evaluation, and human-in-the-loop workflows** across internal and portfolio use cases.
- **Consult as a backend engineer across portfolio companies**, designing distributed services, APIs, data pipelines, integrations, and cloud infrastructure.

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 – June 2026
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.

Open Source & Technical Writing

- **Open source:** Build and maintain LLM infrastructure, agent systems, and research harnesses spanning memory and retrieval, inference gateways, model evaluation, alignment experiments, and model compression.
- **Technical writing:** Write technical essays and interactive explainers on LLM architecture, attention, inference, agents, evaluation, and experimental research at divagr.com.

Technical Skills

AI/ML Systems: LLM Inference, RAG, Post-Training & RL, Model Compression, Long-Context Architectures, World Models, Representation Learning, Agent Orchestration, Multimodal Agents, Alignment, Evals & Red Teaming

Backend & Systems: Distributed Systems, Event-Driven Architectures, Kafka, Redis, PostgreSQL/pgvector, Kubernetes, Docker, CI/CD, Observability, AWS, GCP, Azure

Languages & Frameworks: Python, TypeScript, JavaScript, SQL, Bash, FastAPI, Django, React, Node.js/Express

Education

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