Divyansh Rai

divyanshr988@gmail.com | +91 7999848709 | linkedin.com/in/divyanshrai01 | github.com/Divyansh2202

Summary

AI Engineer skilled in **LLM/VLM fine-tuning & quantization (LoRA, QLoRA, GGUF, INT4)** and **multi-agent, tool-calling system design**. Proficient across the full GenAI stack – RAG pipelines, multimodal intelligence (ASR + NLP + Vision), embedding-based retrieval, low-latency inference optimization – with strong production observability, cost control, and reliability engineering for high-concurrency real-world systems.

Technical Skills

Machine Learning	PyTorch, TensorFlow, Scikit-Learn; Computer Vision, NLP, large-scale training pipelines
Fine-Tuning & Optim.	LoRA, QLoRA, SFT; Unsloth, TRL, LLaMA-Factory; Quantization (GGUF, INT4), llama.cpp, vLLM
Generative AI & LLMs	LLM/VLM system design; RAG pipelines, Prompt Engineering, Multi-Agent & Tool-Calling Systems
LLM Orchestration	Vercel AI SDK, LangGraph, LangChain
Agentic Systems & SDKs	MCP, Gemini SDK, Ollama; Memory: HindSight, Mem0
Backend & Data Infra	Redis, Kafka, PostgreSQL, ClickHouse
Observability	Langfuse, OpenTelemetry; distributed tracing, cost & performance optimization
Vector Databases	pgVector; semantic search, embedding pipelines
Programming	Python (Pandas, NumPy, SciPy), SQL, OpenCV, NLTK, Streamlit, Hugging Face Transformers
Deployment & MLOps	Docker, Git, CI/CD, Edge AI deployment, scalable inference systems
Data & Analytics	Power BI, Matplotlib, Seaborn, SQL

Experience

TapHealth – *AI Engineer, Gurugram, India* *Jan 2025 – Present*

Multimodal Macro-Nutrient Pipeline & GenAI Dietary Optimization Engine *Jan – Jun 2025*

ASR + NLP + Vision, pgVector, ANN, RAG, Behavioral Modeling

- Architected cross-modal pipeline (ASR + NLP + Vision, slot-filling, Gemini + pgVector ANN) over a **10K+ recipe / 30K+ alias** database for daily meal logging, resolving dish matches in **~2s** (DB-backed) to **3–5s** (LLM fallback) at **~90% accuracy**; powered a RAG meal-plan engine with behavioral signals and condition-specific diabetic constraints.

Qwen3-VL-2B Fine-Tuning & Edge Quantization for Food Recognition *Jul – Oct 2025*

LoRA, QLoRA, GGUF, INT4, KV-Cache, Unsloth

- Led end-to-end VLM fine-tuning: **300K+** sample curation, quantity-aware supervision, and edge deployment with **≈68%** compression (6 GB → 1.9 GB, INT4 GGUF) and **≤6s** latency.

AI Coach – Production Conversational Health Agent Platform *Dec 2025 – Present*

Vercel AI SDK, Kafka, HindSight, Redis, ClickHouse, Langfuse, OpenTelemetry, pgVector, Postgres

- Rebuilt AI Coach's agent orchestration layer – replacing an initial LangGraph state machine and dual-layer Mem0 memory with Vercel AI SDK's bounded tool-calling loop and a single HindSight episodic-memory layer – the foundation the meal-logging, diet-plan, and nudge systems below are built on.
- Built a **2-channel** (WhatsApp, Telegram) conversational agent on a bounded (**13-step max**) tool-calling loop with a YAML-defined skill registry, driven by **16+** Kafka-consumed event domains (meal logging, glucose, onboarding) for real-time patient state.
- Built a single-call, **3-tier** meal-logging resolution engine (personal history → pgVector recipe search → LLM fallback) with confidence-gated confirmation and cheap-first alternates sourcing, replacing a legacy multi-round-trip service; fixed production bugs in per-unit macro calculations and hardened LLM prompts against injection.
- Built an LLM-driven diet-plan generator (ClickHouse history + HindSight memory) and an **8-model** behavioral reflection system (identity, routines, barriers, what-works), plus a **4-state** glucose feedback engine closing the log → feedback → plan loop and driving personalized nudges.

IIT Patna (IITP) – *Research Intern, Patna* *June – July 2023*

- Contributed to SERB-funded behavioral biometrics project (deep learning, multilingual handwriting); processed **1200+** pages from **40+** authors for model training.

Projects

MnemOS – Universal Memory Protocol for AI Apps (Open Source) *2026*

FastAPI, pgVector, bge-m3, Ollama, Gemini SDK, MCP, Python SDK, Chrome Extension

- Architected a **2-platform** (ChatGPT, Claude) persistent memory layer – silently extracts facts post-conversation via LLM (dedup at **0.88** cosine similarity), injects relevant context pre-prompt (**0.65** retrieval threshold), stored by user ID for cross-tool portability.
- Built end-to-end: FastAPI server, pgVector semantic store (**1024-dim** cosine similarity), Chrome extension (**5x**-retry offline queue), Python SDK, CLI, a **4-tool** MCP server (remember/recall/forget/list_memories), and web dashboard – fully owned and open-sourced.

Education

B.Tech, Computer Science Engineering *Graduated June 2024 | CGPA: **8.71***

Indore Institute of Science & Technology, RGPV University, Bhopal

Achievements & Certifications

Research Paper – ICICES 2023, IET DAVV *Feb 2023*

"Climate Change Effects on Crop Production: Precision Agriculture via Data Science"

IBM / Coursera: Machine Learning with Python | Data Science Professional Certificate | Tools for Data Science