

Divyasha Agrawal

Delhi, India · +91-8789623763 · divyasha.agrawal@gmail.com · LinkedIn · GitHub

SUMMARY

Senior ML Engineer with 7+ years building and deploying production ML, NLP, and Generative AI systems on Google Cloud. Hands-on with Vertex AI (Gemini, Vector Search, evaluation), agentic orchestration (tool calling, function calling, structured outputs, custom runtimes), production RAG over managed and open-source vector stores, and LLM security (prompt-injection defense, guardrails, policy enforcement). Ships end-to-end on Cloud Run, Docker, Kubernetes, and CI/CD.

EXPERIENCE

AI & Automation Consultant, Wagify (Pet Care)

May 2026 – Present

- Designed the client-acquisition and intake-to-booking workflow and automated it with agentic AI: an LLM agent with tool calling that handles enquiries, qualifies leads, and manages bookings across grooming and dog-walking services.

Senior Data and ML Engineer, Artefact (Nissan & Infiniti)

Apr 2022 – May 2026

- Built API-driven, event-based services on Cloud Run, Cloud Functions, Pub/Sub, and Eventarc, deployed via Docker and Kubernetes, replacing manual workflows with file-triggered automation.
- Owned 15+ production pipelines on a GCP ML data platform (Python, BigQuery, Dataflow, Prefect, dbt) with automated data-quality checks, Cloud Logging alerting, and an on-call rotation serving 1,500+ stakeholders.
- Migrated the platform from legacy VMs to cloud-native with Git, CI/CD (Cloud Build), and Prefect orchestration, making deployments reproducible and auditable.
- Optimized BigQuery across 40+ dashboards (partitioning, clustering, materialized views), cutting compute cost 35% and query latency from 18s to 4s.

Machine Learning Engineer, Rivi Ventures (Travel Product)

Jun 2019 – Apr 2022

- Designed and deployed a RASA conversational AI system (BERT intent classification, custom NER, dialogue management, action server) handling 15K+ daily conversations at 87% accuracy; async FastAPI + Redis backend at 500+ req/s.
- Built recommendation engines for hotels, activities, and trip duration (XGBoost, Random Forest, KMeans) serving 200K+ users, lifting CTR 24%.

Earlier: Project Trainee, BOSCH RBEI (2018–19): CV-based UML-to-code extraction (OpenCV). Student Mentee, Amazon ACMS (2018): VGG16 image recommender, 73% precision on 50K+ images.

GENAI & AUTOMATION PROJECTS

- Investment Brain:** production RAG on Vertex AI (Gemini generation, Vertex AI Vector Search retrieval) behind a model-routing gateway (Gemini, Claude, Ollama, Qwen) with Qdrant/ChromaDB fallback. Automated eval harness (RAGAS-style golden datasets, regression tests) tracking correctness, cost, and latency over daily ingestion.
- The Claw:** agentic system with autonomous tool use and multi-step, stateful orchestration, running a full job-application workflow end to end (Playwright ingestion, SQLite state, local Ollama scoring, Claude API generation) with no manual steps.
- BiteBuddy:** agentic CLI on MCP with tool calling, structured outputs, and decision-making, hardened with guardrails and policy enforcement (input validation, prompt-injection defense, access control on tool invocation).
- Kakooshah:** multimodal WhatsApp assistant with Gemini vision-based product matching, pgvector retrieval over 13K+ SKUs, async FastAPI on Railway.

TECHNICAL SKILLS

Google Cloud AI: Vertex AI (Gemini, Vector Search, evaluation), Cloud Run, Cloud Functions, Pub/Sub, Eventarc, BigQuery, Cloud Logging

Agentic & LLM Ops: agent orchestration (autonomous tool use, multi-step / stateful workflows, custom runtimes), LangChain, ADK-adjacent, MCP, RAG, embeddings and retrieval, model routing, LLM evaluation (RAGAS-style, golden datasets, regression tests), guardrails, prompt-injection defense, access control, prompt engineering

LLMs & Vector Stores: Google Gemini, Anthropic Claude, open-source (Ollama, LLaMA, Qwen); Vertex AI Vector Search, pgvector, Qdrant, ChromaDB, FAISS

Infra & Platforms: Docker, Kubernetes, GCP, AWS (S3, Lambda), CI/CD (Cloud Build), Git, Terraform, FastAPI (async), REST APIs, Prefect

Data & ML: BigQuery, dbt, ELT/ETL, data quality, identity resolution; transformer fine-tuning (BERT, NER), PyTorch, TensorFlow, Scikit-learn, XGBoost

Languages: Python, SQL, C/C++; Databricks (Certified Professional)

CERTIFICATIONS & EDUCATION

Certifications: Databricks Certified Data Engineer Professional · Databricks Certified ML Professional

B.Tech, Computer Science, Vellore Institute of Technology (CGPA: 8.57 / 10)

Jul 2015 – May 2019