

Umarfarook Gurramkonda

Machine Learning Engineer

Bangalore, India | +91 8897574189 | umarfarook0yt@gmail.com | umarfarook-ai.vercel.app
github.com/Umarfarook1 | linkedin.com/in/umarfarook-gurramkonda

SUMMARY

Machine Learning Engineer (founding ML hire) at an early-stage AI startup, shipping production LLM systems on Google Cloud: multi-stage agent orchestration, Retrieval-Augmented Generation (RAG), NL-to-SQL over BigQuery, and LLM evaluation harnesses. Open-sourced an MCP (Model Context Protocol) BigQuery server on PyPI. Hands-on across machine learning, deep learning, Natural Language Processing (NLP) and computer vision, from data and features through training, evaluation and deployment (Docker, CI/CD). Strong Python (FastAPI, PyTorch).

EXPERIENCE

Founding ML Engineer | Hypeon AI | Bangalore, India

Oct 2025 - Present

- Architected and shipped a production multi-stage LLM orchestration service with hierarchical agentic workflows (routing, retrieval, analysis, composition sub-agents), SSE streaming, and Pydantic-validated structured outputs, powering a live competitive-intelligence product across 4 ad platforms (Meta, Google, TikTok, Pinterest).
- Designed and deployed a natural-language-to-SQL system over a BigQuery data warehouse with schema introspection, synonym matching, and dry-run cost guardrails (Claude Haiku primary, Gemini fallback).
- Shipped a remote MCP (Model Context Protocol) server exposing 12 analysis tools with OAuth 2.1 + PKCE auth, plan-gated quotas, and cross-pod request cancellation over Redis pub/sub.
- Integrated Claude, Gemini, and OpenAI behind a unified fallback router with structured-output validation and injection guardrails; deployed on GCP Cloud Run (Cloud SQL/PostgreSQL, Redis, Prometheus, GitHub Actions CI/CD).

Freelance ML / AI Engineer | Self-employed | Bangalore, India

Oct 2024 - Sep 2025

- Built an AI-powered inventory management system for a Bangalore retail client: automated invoice generation via LLM-based extraction (OpenAI API), demand forecasting with scikit-learn, real-time stock alerts, and a visualization dashboard.

Backend Developer Intern | Synclovis Systems | Bangalore, India

Jun 2024 - Sep 2024

- Built RESTful backend services (Node.js, Express, MySQL) and contributed to an internal LLM-based healthcare assistant: RAG retrieval (LangChain + FAISS) over clinical documents with guardrails to reduce hallucinations.

EDUCATION

B.Tech, Computer Science | K.S.R.M College of Engineering, JNTU Anantapur

2020 - 2024 (CGPA: 8.14 / 10)

PROJECTS

mcp-bigquery-evals - BigQuery MCP Server with Cost Guardrails and NL-to-SQL Eval Harness

Python, MCP, BigQuery, pytest

github.com/Umarfarook1/mcp-bigquery-evals | PyPI: pypi.org/project/mcp-bigquery-evals

- Published an open-source MCP server (PyPI, uvx-runable) exposing 7 read-only BigQuery tools with mandatory dry-run cost caps and 7 stable structured error codes that let LLM agents self-correct instead of crashing.
- Enforced read-only access at the statement level from the dry-run statement type, so DDL is refused before the byte cap it would otherwise pass; a fake BigQuery client enables 211 credential-free unit tests in CI (strict mypy, ruff).

TrustBench - Production-Readiness Evaluation Harness for AI Support Agents

Python, Gemini, LLM-as-judge

github.com/Umarfarook1/trustbench

- Built an eval harness scoring a tool-using RAG agent against versioned golden sets on 6 trust dimensions, combining deterministic checks with temperature-0 LLM-judge rubrics.
- Implemented per-intent regression detection with exact McNemar tests on paired pass/fail outcomes, surfacing category-level drops that flat aggregates hide; 82 offline unit tests via fake LLM clients, green CI.

Nano-LLM-from-scratch - GPT-2 Rebuilt with Modern Components

Python, PyTorch

github.com/Umarfarook1/Nano-LLM-from-scratch

- Implemented a decoder-only transformer from scratch with RoPE, RMSNorm, SwiGLU, a KV-cache and a byte-level BPE tokenizer; 257 offline tests including cached-versus-full-forward logit equivalence.
- Trained on CPU for 2,500 steps with validation loss falling 4.94 to 2.90, and published the log, the benchmark and the sample output alongside the claims rather than the numbers alone.

street-view-plate-blurring - YOLOv8 License-Plate Detection and Privacy Redaction

Python, PyTorch, YOLOv8, OpenCV

github.com/Umarfarook1/street-view-plate-blurring

- Trained a YOLOv8n object detection model to mAP@0.5 0.78 (precision 0.83, recall 0.74) at ~4 ms/image on a Tesla T4, grounding preprocessing choices in a data audit and EDA of 8,740 labeled boxes.
- Built a recall-first privacy redaction pipeline (box-scaled Gaussian blur) that expands every detection before blurring, trading sharpness around the plate for a lower miss rate.

Cargo Concierge - Agentic Freight-Quoting Copilot

TypeScript, Next.js, Mastra, PostgreSQL

github.com/Umarfarook1/Cargo-Concierge | Live: cargo-concierge.vercel.app

- Built an agentic copilot converting free-form freight quote requests into ranked airline options and a draft reply via a 5-stage streamed pipeline.
- Hand-labelled a 15-item extraction eval scoring 14/15 exact match, with prompt ablations showing that removing the rules block costs about 33 percentage points of accuracy.

ipl-data-analysis - Match-Outcome Modeling on 1,095 IPL Matches

Python, scikit-learn, XGBoost

github.com/Umarfarook1/ipl-data-analysis

- Engineered chronological venue-history features and benchmarked 6 classifiers (logistic regression with SMOTE, Random Forest, AdaBoost, XGBoost, soft-voting ensemble) on a 414-feature matrix; the best reached ROC-AUC 0.547, stated plainly rather than inflated.

More on GitHub (github.com/Umarfarook1): Tiny-diffusion (DDPM with classifier-free guidance and a DDIM sampler, from scratch, 111 tests) | rag-document-qa (RAG with a retriever eval harness) | youtube-shorts-performance-prediction (a negative result and its leakage forensics).

SKILLS

Programming: Python (Pandas, NumPy), SQL, TypeScript / JavaScript

Machine Learning: scikit-learn, XGBoost, feature engineering, cross-validation, clustering (K-Means, PCA), class imbalance (SMOTE), statistical testing

Deep Learning / NLP: PyTorch, Hugging Face Transformers, sentence-transformers (BERT-family), Natural Language Processing (NLP), embeddings, transformer internals (RoPE, RMSNorm, SwiGLU, KV-cache), diffusion models (DDPM)

Computer Vision: YOLOv8 (Ultralytics), OpenCV, object detection, image preprocessing

LLMs / GenAI: RAG, agentic systems, multi-stage orchestration, LLM evaluation (eval harnesses, LLM-as-judge, golden sets), Model Context Protocol (MCP), structured outputs (Pydantic), streaming (SSE); Claude, Gemini, OpenAI GPT; LangChain; FAISS, chunking, reranking

Backend / Cloud: FastAPI, REST APIs, async SQLAlchemy, Node.js; GCP (Cloud Run, Cloud SQL, BigQuery), AWS (S3, EC2), PostgreSQL, Redis

MLOps / DevOps: Docker, Git, GitHub Actions CI/CD, model deployment and monitoring (Prometheus), pytest, mypy, ruff