Ghanshyam Paunikar

ghanshyampaunikar.com | ghanshyampaunikar@gmail.com | linkedin.com/in/ghanshyampaunikar | github.com/GhanshyamPaunikar |
+91 7387123317 | Pune, India

Full-Stack AI, RAG, LLM Evaluation, AI Agents, LangGraph, Databricks / PySpark, PostgreSQL / pgvector, Azure AI Foundry, React

PROFESSIONAL SUMMARY

AI Engineer with over three years of experience building LLM-powered applications end to end, across UI, backend services, and model orchestration, and an MSc in Data Science and AI from the University of Liverpool. I design RAG pipelines for enterprise knowledge retrieval and grounding, build LLM workflows and agents with multi-step reasoning and tool use in LangChain, LangGraph, and AutoGen, and integrate models through Azure AI Foundry and hosted APIs. I specialise in LLM evaluation: frameworks for accuracy, hallucination detection, relevance, and safety, automated benchmarking of prompt chains and agents, and metrics such as precision, recall, and grounding fidelity. I build data pipelines with Databricks, PySpark, and Delta Lake, support feature stores and knowledge bases for AI, retrieval layers with PostgreSQL and pgvector, Python APIs, and React front ends, ready for production with logging, monitoring, and security.

PROFESSIONAL EXPERIENCE

LLM Engineer
April 2024 – September 2026
Sigma AI
London, UK

- Built LLM-powered applications end to end, from backend services and model orchestration to the review tools used daily by 300+ transcribers and reviewers.
- Developed LLM evaluation frameworks with golden datasets scoring accuracy, groundedness, faithfulness, and hallucinations, plus automated regression benchmarking of prompt chains before every release.
- Defined and monitored LLM performance metrics with Langfuse and OpenTelemetry, and ran experiments comparing models and prompts, including fine-tuning with LoRA/PEFT.
- Built LLM workflows and agent-style chains with LangChain and LangGraph using multi-step reasoning, tool use, and retrieval for grounding.
- Built data pipelines and Python APIs integrated with internal data sources, decreasing analysis turnaround by 30%, with logging, monitoring, and data privacy controls.
- Worked with platform, data, and operations teams to ship production-ready AI, reducing error rates by 12%.
- **Technologies:** Python, LangChain, LangGraph, LLM APIs, LoRA/PEFT, Pydantic, pytest, FastAPI, Docker, Langfuse, OpenTelemetry, SQL

Data Scientist Intern
October 2023 – March 2024
Virtual Engineering Centre, University of Liverpool
Liverpool, UK

- Built an LLM application end to end in Python with the OpenAI API: form-based input, prompt orchestration, and automated outputs, cutting manual processing time by approximately 60%.
- Embedded generative AI into an existing business workflow used for real case-study submissions.
- Evaluated results against the manual process and reported performance through Power BI dashboards.
- Developed Python-driven automation workflows within NVIDIA Omniverse, reducing repetitive 3D layout design effort by approximately 40%.
- **Technologies:** Python, OpenAI API, SQL, Power BI, TensorFlow, NVIDIA Omniverse

Data Scientist
February 2021 – October 2021
Zummit Infolabs
Bangalore, India

- Developed and deployed a PyTorch UNet model for MRI segmentation, improving cancer detection accuracy by 15%.
- Built data preprocessing and deployment pipelines, reducing training time by 20%.
- Communicated findings to non-technical stakeholders weekly.
- **Technologies:** Python, PyTorch, NumPy, Pandas, OpenCV

PROJECTS

Innerbloom: 100% local AI journal with zero telemetry. Streaming agent chat across three personas (Companion, Observer, Challenger), 16-emotion tagging, a Wellbeing Radar for burnout and spiral patterns, a People social CRM, RAG over past entries, and a Memory Graph.
Technologies: *Ollama, FastAPI, SQLite, LangGraph*

F1 Race Predictor: Gradient-boosting ensemble trained on 8 seasons of F1 data (3,458 samples, 12 engineered features) with 1,000x Monte Carlo simulation for win probabilities; CV MAE 3.41 positions. Live with podium and full 20-driver grid predictions.
Technologies: *XGBoost, HistGBR, FastAPI, Vue.js, Railway*

Zeul: Android Wallpaper App: Open-source Android wallpaper app with HD and AMOLED feeds, community uploads, and an auto wallpaper changer, built by two developers in about two weeks with Claude Code agents. I built the Android app: UI, data layer, preview, search, onboarding, and background work.
Technologies: *Kotlin, Jetpack Compose, Claude Code agents*

CarbonWatcher: Published pip package (pip install carbonwatcher) that tracks energy, CO2, and cost for ML training runs and LLM API calls (Anthropic, OpenAI, Ollama, Hugging Face), with location-aware grid intensity and team dashboards.
Technologies: *Python, Intel RAPL, NVIDIA NVML, pytest, GitHub Actions*

AI NPC: Living Characters: MSc dissertation. NPCs with real-time voice, lip-sync and facial animation, and emotional responses; characters hold context-aware conversations and adapt to player emotions. 45% engagement lift.

Technologies: *GPT-3.5, Unity, C#, ElevenLabs, Meta Oculus Voice*

SpaceX Falcon 9 Landing Prediction: IBM Data Science capstone predicting whether the Falcon 9 first stage lands successfully, to estimate launch costs, using classification models, EDA, SQL, interactive maps, and a Plotly Dash dashboard.
Technologies: *Python, SQL, Scikit-learn, Folium, Plotly Dash*

ASL Sign Language Recogniser: Real-time ASL alphabet recogniser that tracks 21 hand landmarks per frame and maps their geometry to 26 letter signs with a custom classifier, showing live confidence scores from webcam.
Technologies: *PyTorch, MediaPipe, OpenCV*

Image Resolution Upscaler: Prompt-guided 4x super-resolution via latent diffusion, generating sharp high-frequency detail; runs locally with a Gradio UI and a cached 1.7 GB model.
Technologies: *Stable Diffusion x4, PyTorch, diffusers, Gradio*

Piano Tutor: Desktop and web app to learn piano with falling notes and every key named; practice mode waits for the right note, with separate hands, adjustable tempo and loops, and scrolling sheet music. Works with any MIDI keyboard on Mac, Windows, and web.
Technologies: *Web MIDI, Python tooling*

TECHNICAL SKILLS

AI Applications & Agents: RAG, Grounding, LLM Workflows, AI Agents (Multi-Step Reasoning, Tool Use), LangChain, LangGraph, AutoGen, CrewAI, MCP, Summarisation, Search

LLM Evaluation: Evaluation Frameworks (Accuracy, Hallucination, Relevance, Safety), Precision, Recall, Grounding Fidelity, Golden Datasets, Ragas, DeepEval, Automated Benchmarking, Prompt Engineering

Data & Retrieval: Databricks, PySpark, Delta Lake, Feature Stores, PostgreSQL, pgvector, Embeddings, Indexing Strategies, Vector Search, Snowflake, SQL, Data Modelling

Azure & Platforms: Azure AI Foundry, Azure OpenAI, Azure AI Search, Microsoft Copilot Studio, AWS, GCP

Full Stack: Python, FastAPI, Flask, REST APIs, React, TypeScript, Streamlit, Vue.js

Production Readiness: Logging, Monitoring (Langfuse, OpenTelemetry), Docker, Kubernetes, CI/CD (GitHub Actions, Jenkins), Security, Entra ID, OAuth

ML: PyTorch, Hugging Face, Fine-Tuning (LoRA, QLoRA), Scikit-learn, XGBoost

EDUCATION

MSc Data Science and Artificial Intelligence — *September 2022 – September 2023*
University of Liverpool — *Liverpool, UK*

B.E. Mechanical Engineering — *2016 – 2020*
Raisoni College, RTMNU (Rashtrasant Tukadoji Maharaj Nagpur University) — *Nagpur, India*

PUBLICATIONS & ARTICLES

Innerbloom AI Journal: What I Built After Months of Journaling Into a Void
Fine-tuning Falcon-7B with QLoRA on a Mental Health Conversational Dataset
Modern AI Applications in Video Gaming
Counting Brine Shrimp in Videos with YOLOv5 and Active Learning

CERTIFICATIONS

Claude Code 101	*Anthropic*
Data Science Professional Certificate	*IBM*
Generative AI: Elevate Your Data Science Career	*IBM*
Machine Learning with Python	*IBM*
Responsible Generative AI Specialisation	*University of Michigan*
Elements of AI	*University of Helsinki*
Artificial Intelligence	*DRDO, Govt. of India*
Data Visualization	*Kaggle*
ML, Data Science & Deep Learning with Python	*Udemy*
Python Programming for Artificial Intelligence	*Udemy*
Artificial Intelligence & AR in Unity + Xcode	*Udemy*
Supervised ML: Logistic Regression & Naïve Bayes	*Great Learning*
Unsupervised Machine Learning: K-Means	*Great Learning*
Machine Learning with Python: Foundations	*LinkedIn Learning*
SQL for Data Analysis	*LinkedIn Learning*

LANGUAGES

English: Fluent | **Hindi:** Native | **Marathi:** Fluent | **German:** A1

VOLUNTEERING

Spread Smile Foundation — *India*
Volunteered with community outreach programmes supporting education, healthcare, and community development.