Rishabh Hota

AI/ML Researcher & Engineer
rishabhhota19@gmail.com | LinkedIn

PROFESSIONAL SUMMARY

AI professional with 3+ years of experience translating cutting-edge AI research into production-ready prototypes and AWS-hosted pipelines. Proven leader of cross-functional teams in healthcare and ed-tech, with strong expertise in deep learning, computer vision, generative AI and model architecture, fine-tuning LLMs, multi-agent systems, harness engineering, loop engineering, Model Context Protocol (MCP), and scalable architectures.

WORK EXPERIENCE

Sr. AI Researcher — Simplilearn (On-site, Bengaluru, Karnataka) *Mar 2026 – Present*

- Automating complex workflows
- Open source
- Agentic orchestration
- Research on new tech (hands-on)

AI Researcher — Simplilearn (On-site, Bengaluru, Karnataka) *Mar 2025 – Mar 2026*

- Track and dissect emerging AI breakthroughs (MCP, agentic AI, LLMs, Gemini) into hands-on prototypes.
- Build, validate, and document research demos; curate a code-backed learning hub.
- Design bite-sized, code-driven modules; influence roadmap for next-gen AI features.
- Build scalable GenAI applications around the research.

AI Engineer — OsteosConnect INC (Remote, California) *Feb 2024 – Mar 2025*

- Led clinicians and engineers to deliver a 96%-accurate osteoarthritis classification pipeline (X-ray → S3 → Lambda → EC2 → Prediction → Lambda → EMR).
- Surpassed competitor RBKnee and achieved >90% concordance with radiologists in SRM Trichy trials.
- Implemented a continuous feedback loop for model retraining using doctor-validated ground truth.

AI/ML Intern — OsteosConnect INC (Remote, California) *Jul 2023 – Jan 2024*

- Hypothesized and prototyped ML approaches for automated OA grading.
- Conducted EDA on DICOM imagery; established baseline classifiers (logistic regression, Random Forest).

Data Science & ML Intern — Smartknower Pvt Ltd (Bangalore) *Nov 2020 – Jan 2021*

- Executed full ETL pipeline, feature engineering, and predictive modeling.
- Developed Power BI dashboards for sales forecasting; presented actionable insights.

EDUCATION

B.E., Electronics & Communication — B.M.S. College of Engineering, Bangalore | GPA: 7.65/10 *Aug 2019 – Aug 2023*

Intermediate (PCM) — Shiv Jyoti Convent School, Kota *Apr 2017 – May 2018 | 80%*

Matriculation — Sri Chaitanya Techno School, Vizag *Apr 2015 – May 2016 | CGPA: 10*

KEY SKILLS

Forward Deployed Engineering · System Design · Open Source · Azure AI Foundry · AWS Bedrock · Agent & LLM Evaluation · Agent & LLM Guardrailing and Deployment · Token Optimization · Agentic AI and Harness · Automation Engineering · Chatbot Design · AI & Agentic AI Orchestration · LLM Architecture and Fine-tuning · RAG at Production · MCP · Healthcare Automation · Research Publication · Claude Code, Claude Science · Claude Skills · Speech Generation · Classification · GenAI Applications

TECHNICAL SKILLS

AI & ML: Generative AI · Agentic AI and Agentic RAG · Multi-Agent Systems · Retrieval-Augmented Generation (RAG & Graph RAG) · Model Context Protocol (MCP) · Chatbots · Vector DBs (Redis, FAISS, etc.) · Machine Learning · Deep Learning · Computer Vision · CNNs · NLP · Transformers · Segmentation Models · X-ray Annotation

Languages & Frameworks: Python · Scikit-learn · TensorFlow · PyTorch · PyTorch Lightning · LangChain · LangGraph · Google ADK · Google Gemini APIs · OpenAI SDK · CrewAI · Pandas · NumPy · Asyncio · FastAPI · Flask · FastMCP · gRPC · Pydantic · PyDICOM

LLM Fine-tuning & SLMs: Unsloth · LoRA & QLoRA (PEFT) · Supervised Fine-Tuning (SFT) · Hugging Face Transformers & TRL · 4-bit / 8-bit Quantized Fine-tuning · Fine-tuning SLMs (Gemma, Llama, Phi)

LLM Inferencing & Serving: vLLM · Ollama · llama.cpp · Model Quantization (GGUF, AWQ) · KV-Cache Optimization · Batching & Streaming Inference · OpenAI-Compatible API Serving

Cloud & DevOps: AWS (S3, Lambda, SageMaker, EC2, Bedrock) · Docker · Azure AI

Data & Analytics: Data Science · Predictive Modeling · EDA · ETL · DICOM Processing

Foundational CS: Algorithms · Data Structures (DSA) · Object-Oriented Programming (OOP) · DBMS · Computer Networks (CN) · System Design for AI

PROJECTS

AI Teddy Bear (MVP)

- Built an MVP AI teddy bear that narrates stories to children with a unique character voice and style. All-natural storytelling with personalized stories to feel more comfortable (can't disclose much).
- Open source; hardware- and app-ready; guardrailed and secure for children's use.

Ed-Tech WebApp

- Built an ed-tech platform where learners explore courses and join cohorts, with automated cohort and instructor mapping, automated communication channels — agent posts (WhatsApp) — and automated payroll.
- Stack: frontend and middleware — Vercel · error tracking — Sentry · web analytics — Umami · database — Supabase · payments — Razorpay · in-house revenue analytics.
- Enhanced admin dashboard: all analytics, cohort creation, instructor pool, payroll, live error tracking.

Claude Personal Health Doctor

- Made a Doctor Skill containing medical resources, links for the agent to visit later for any query, guidance on how a doctor should behave and reason, and a strict workflow to investigate any evidence with current information.
- Kept all my medical documents segregated in folders by date; used Graphify to make a second brain for Claude — processed the docs into nodes (communities, colonies, etc.). Query and be healthy.

Collaterals Agent (On-site tool development) *Simplilearn · Jun–Aug 2025*

- Legacy process required 2 teams (4 members each) working 7–10 days to create SD/SL pages, brochures, and 7 emailer types per course launch; the new N8N-based agentic system reduced turnaround to just 1 day, including reviews and approvals.
- Built with a fine-tuned GPT model with deep course knowledge to draft accurate, relevant marketing content automatically.
- Architecture: 1 Head Agent coordinates 2 sub-agents — a Researcher (gathers and validates course information via SerpAPI) and an Analyst (checks content structure and data accuracy, extracts the Table of Contents).
- Integrated with Google Docs for automatic document creation and real-time updates throughout the workflow.

Google AI Edge Gallery: Gemma 3n Android-compatible (On-site research) *Simplilearn · May–Jun 2025*

- Tested the AI Edge Gallery in Android Studio and understood the project structure.
- Hosted Android-compatible Gemma 3n in a custom chatbot app; understood the power of offline intelligence.

TurboTrade (Zerodha MCP) (On-site project) *Simplilearn · Apr–May 2025*

- TurboTrader is a lightning AI copilot that lets you research, plan, and trade Indian markets in one chat: LangChain handles orchestration, Groq delivers ultra-fast reasoning, and Zerodha Kite MCP streams secure real-time account data.
- Ask: "Show my live P&L, scan for bullish breakouts, then sell half my HDFC Bank." The agent fetches positions through MCP, blends fresh news with LangChain RAG, and Groq streams insights at human speed — with one-tap options to buy, sell, set stops, or queue deeper research.
- MCP keeps keys out of prompts and uses OAuth scopes to guard capital; all steps are logged and reversible. Swap models, add brokers, or plug in custom research feeds without rewrites — compressing research, decision, and execution into a single dialogue while staying compliant with Indian rules.

Scalable Sales Training Chatbot (FAISS, Gemini, Google ADK, Streamlit) (On-site project) *Simplilearn · Apr–May 2025*

- Designed a high-throughput sales-enablement chatbot: vector-indexed knowledge base with FAISS, LLM (Gemini) orchestration via Google ADK.
- Developed a Streamlit interface for interactive training scenarios and real-time feedback; engineered for horizontal scalability.

AI-Powered OA Classification & Reporting (On-site project) *OsteosConnect · Jan 2024 – Sep 2024*

- Automated end-to-end flow: X-ray capture on React/Android → encrypted S3 upload → SageMaker inference → web-based report.
- Closed feedback loop to retrain models with doctor-validated labels.

COVID-19 Decision Support System *Team of Four · Oct 2022 – Apr 2023*

- IoT-enabled home assessment combining sensor data, symptom checklists, and X-ray/CT analysis via DL + rule-based logic.
- Designed a two-stage, rule-based diagnostic pipeline integrating ML models trained on patient vitals, symptoms, and chest X-rays: stage one performs a preliminary assessment from vitals and symptoms; stage two adds chest X-ray analysis for a deeper diagnosis.
- Developed an intuitive UI guiding patients through the workflow: Phase 1 asks screening questions; if further evaluation is indicated, Phase 2 prompts a chest X-ray upload, and a final rule-based engine synthesizes all inputs into a comprehensive report with recommended next steps.
- Awarded 6th place college-wide.

News Search & Research Tool (GenAI) (Personal) *Jan 2024 – Feb 2024*

- Built pipeline: NLP-driven query parsing → web scraping → corpus assembly → RAG → prompt-based Q&A interface.

RESEARCH & PUBLICATIONS

- *Do Quantum-Inspired Feature Maps Help Vision Transformer Attention? A Matched-Budget Dequantization Study* — communicated to QMI (Springer).
- *QFPViT: A Fidelity-Phase Quantum Self-Attention Operator for Vision Transformers — Design and Evaluation on MedMNIST* — communicated to Quantum Journal.
- *Multi-Modal Decision Support for COVID-19 Prediction* — communicated to Inderscience.
- *LCSNet: Lightweight Caries Segmentation Network for dental caries segmentation from smartphone photographs* — published in Intelligence-Based Medicine (Elsevier, Impact Factor 2.3), Oct 9, 2025. Built a dataset of smartphone images with varying resolution & illumination and proposed a novel lightweight LCSNet architecture achieving F1-score 71.7% and IoU 56.8% with significantly fewer parameters and FLOPs; results validated by expert dentists. View Paper (ScienceDirect)
- *COVID-19 Classification with Chest CT* — submitted to Intl. Journal of Computing Science & Mathematics (Inderscience, Scopus & ESCI). Hota R.K., Akhilesh S. View Paper (IEEE Xplore)
- *Depression Identification Decision Support System (DIDSS)* — in Springer: Advanced Machine Learning for Real World Applications. Hota R.K., Verma N. View Paper (Springer)
- *OA Classification and Reporting using Agentic AI and Deep Neural Networks* — ongoing with OsteosConnect.

CERTIFICATIONS & TRAINING

LLM Fine-tuning (DeepLearning.AI) · Fundamentals of Generative AI (Databricks, 2024) · NVIDIA Deep Learning Institute: Vision & Multi-Modal (2020) · Python 3 Programming Specialization (Univ. of Michigan, Coursera) · Practical Deep Learning (Intel, Coursera) · Data Science for All (E&ICT, NIT Warangal) · AI & Deep Learning (E&ICT, IIITDM) · ML Foundations: A Case Study Approach (Univ. of Washington, Coursera) · Machine Learning for All (Univ. of London, Coursera)

ACHIEVEMENTS

- With zero knowledge, started with a goal to build a better attention mechanism using quantum algorithms — successfully created QFPViT and submitted the paper to Quantum Journal.
- Won the Best Team Award at Simplilearn for Q3; automated collateral creation for any new program launch, shrinking the time from 7–10 days to just 1 day.
- Published my novel architecture in Elsevier (Intelligence-Based Medicine, Impact Factor 2.3).
- Built a scalable sales training system (Simplilearn).
- Led MVP development with doctors & researchers at OsteosConnect; my osteoarthritis classification and reporting solution outperformed competitors (RBKnee).
- 6th prize, major project showcase (4th year, BMSCE).
- Published in Springer & Inderscience journals.
- 2nd prize, COCA-COLA Cup district-level football.
- Poetry published in Amar Ujala (Dainik Bhaskar).

SOFT SKILLS & INTERESTS

Soft Skills: Team Management & Leadership · Teamwork · Problem-Solving · Communication · Curiosity

Interests: AI Research · Competitive Football · Professional Poetry