

Om Rajesh Pawar

AI Solutions Engineer | Generative AI | Customer-Facing Solutions

Dubai, UAE | +971 50 207 5624 | ompawarcloud@gmail.com | UAE Driving License Holder

linkedin.com/in/om-pawar-rajesh | github.com/omrajeshpawar | ompawar.com

Professional Summary

AI Solutions Engineer and AWS Certified Solutions Architect building Generative AI and agentic systems that solve operational business problems. Experienced in translating customer requirements into practical AI solutions, leading technical discovery and product demonstrations, and delivering measurable improvements in efficiency, accuracy, cost, and reliability across enterprise workflows. Strong foundation in AWS, Generative AI, RAG, LLM orchestration, computer vision, and cloud architecture, with a focus on translating AI capabilities into practical, measurable business value.

Education

Master of Science (MSc), Artificial Intelligence and Machine Learning Sep 2025 - Sep 2026
University of Birmingham Dubai, UAE | On track for distinction

Bachelor of Engineering (BE), Artificial Intelligence and Data Science Jun 2021 - Jun 2025
PVG's College of Engineering and Technology (SPPU), Pune, India | CGPA 8.77 out of 10.0

Professional Experience

AI Cloud Consultant | BidInspect, HCF Tech Services Pvt. Ltd, Dubai, UAE Dec 2025 - Sept 2026
Tools: AWS Bedrock, AWS, Vertex AI, GCP, Python, RAG, Agentic AI, Explainable AI, LLM Orchestration, Workflow Automation

- Built an AI evaluation platform that **reduced tender review time by 90%**, validated across **50+ tenders and 200+ bidder submissions**.
- Scoped AI solutions and POCs** with procurement stakeholders, translating business requirements and workflow pain points into technical solution designs.
- Reduced hallucination rate **to 1% across 100+ scanned technical documents** using citation-backed, schema-constrained outputs.
- Maintained **zero critical data loss across 1,000+ extracted fields** through human-in-the-loop verification of ambiguous and conflicting evidence.
- Delivered **10+ live product demonstrations** for non-technical procurement stakeholders.
- Prepared technical proposals and participated in commercial discussions with prospective customers, communicating solution architecture, scope, implementation approach, capabilities, and expected business value.
- Incorporated customer feedback into product changes**, iterating workflows and AI functionality based on real stakeholder requirements.

Cloud Engineer Intern | Verve Global Services, Pune, India Jun 2025 - Sep 2025
Tools: AWS Lambda, Amazon S3, CloudFront, Route 53, EC2, CloudWatch, QuickSight, IAM, WAF

- Re-architected video platform infrastructure for high availability and regional failover using Lambda, S3, CloudFront, Route 53, and EC2 across 3 AWS availability zones.
- Reduced content-delivery and access risk with signed URLs, origin-access controls, and routing fixes.
- Built CloudWatch and QuickSight dashboards that identified over **10% in recurring operational cloud spend savings**.
- Hardened security posture with least-privilege IAM roles, encryption, security groups, and WAF rules.
- Presented cost, reliability, and security trade-offs directly to management and engineering stakeholders.

AI Data Engineer Intern | PVG's College of Engineering and Technology, Pune, India Jan 2025 - Mar 2025
Tools: Python, OpenCV, MediaPipe, TensorFlow, Llama 3, RAG, React, Node.js, Flask, MongoDB

- Built a pose-estimation system reaching **approximately 92% accuracy** across 50+ postures using OpenCV, MediaPipe, and joint-angle logic.
- Gathered requirements directly from national-level yoga practitioners to define the coaching workflow.
- Integrated a RAG pipeline that generated personalized routines from a curated pose library using Groq LLM.
- Shipped a full-stack web application (React, Node.js, Flask, MongoDB) supporting real-time video inference at 90 FPS.

Projects

AI-Powered Inventory Management System | Independent Project

May 2026 - Present

Tools: AWS Bedrock, Claude Agent SDK, Computer Vision, Python, PIL

- Identified warehouse inventory counting as a labor-intensive workflow and designed a GenAI/CV solution on AWS Bedrock to automate counting across diverse item categories and types.
- Reached **98.8%** bin ID extraction accuracy at approximately **2.4s latency** and **\$1.83 per 1,000 images**.
- Built a 5-way model voting ensemble that reached **94.7%** item counting accuracy.
- Designed a capture-quality overlay that detects glare and motion blur and blocks bad images before they reach the model which prevents a wasted inference call.
- Modeled a labor-saving opportunity exceeding **100 hours per week** in labor savings for large-scale warehouse operations.

Adaptive Routing for Intelligent AI | MSc Thesis

Mar 2026 - Sept 2026

Tools: Python, Ollama, Local LLM Inference, Hybrid AI Serving, Predictive Routing, Telemetry

- Built a local-first AI gateway that routes requests between local inference and cloud overflow using live queue depth and latency telemetry.
- **Reduced SLA violations from 35% to 6%** at a matched cloud-offload budget using a calibrated risk model (0.979 ROC-AUC).
- Diagnosed a **10x GPU thermal slowdown** at 80 degrees Celsius under sustained load, then added cooldown controls and automated tests to keep the pipeline reliable.

Self-Healing AI Immune System | Cybersecurity Lead

Jan 2026 - Mar 2026

Tools: AWS Bedrock, Cybersecurity, LangChain, Python, Claude, Cost & Latency Benchmarking, Prompt Engineering

- Built a four-layer AI security gateway combining heuristic filtering, vector similarity search, LLM-based intent analysis, and policy auditing.
- Implemented regex-based detection for SQL injection, XSS, Base64 payloads, and prompt override attempts.
- Implemented self-healing threat memory that stores newly detected attack vectors for future detection without manual rule updates.
- Built semantic detection using Chroma vector databases, HuggingFace embeddings, and Groq-hosted Llama 3 through LangChain.

Technical Skills

Generative AI & Agentic Systems: Generative AI, Agentic AI, Retrieval-Augmented Generation (RAG), LLM Orchestration, Prompt Engineering, Model Evaluation, Human-in-the-Loop AI, Explainable AI, Document Intelligence, OCR, Claude Code, Codex, LangChain, Pydantic

AI and Machine Learning: TensorFlow, Computer Vision, OpenCV, MediaPipe, Local LLM Inference, Ollama, Hugging Face, ChromaDB

Cloud and AI Platforms: AWS, Amazon Bedrock, Google Cloud Platform, Vertex AI, Gemini, Serverless Architecture, Cloud Security, Disaster Recovery, High Availability, Cost Optimization, Observability

Software and Data: Python, SQL, Pandas, NumPy, REST APIs, Flask, Node.js, React, MongoDB, Supabase, Git, Power BI, Amazon QuickSight

Solution Engineering: Technical Discovery, Requirements Gathering, Solution Scoping, POC Planning, AI Solution Design, Technical Proposals, Product Demonstrations, Commercial Discussions, Stakeholder Communication

Certifications

- AWS Certified Solutions Architect - Associate (SAA-C03), Amazon Web Services ([click here to verify credential](#))
- Claude Academy: Model Context Protocol: Advanced Topics, Anthropic ([click here to verify credential](#))
- Claude Academy: Building with the Claude API, Anthropic ([click here to verify credential](#))
- Generative AI with Diffusion Models, NVIDIA
- Google Cloud Foundations, GCP
- Fundamentals of Deep Learning, NVIDIA