Mangesh Tanaji Kashid

Forward Deployed Engineer / Applied AI Engineer

+91 76660 55989 | manguss.kashid@gmail.com | Bangalore

linkedin.com/in/mangeshkashid | github.com/LEO2822 | eemky.xyz

Summary

Forward Deployed Engineer / Applied AI Engineer with 5+ years of experience, including internships, building and delivering client-facing AI and software solutions. Architected and built SensAI, an enterprise AI Agent Studio adopted by 50+ internal users, delivering 6 production agents and reducing manual effort by over 60%. Own delivery from discovery and technical scoping through implementation and production deployment, translating ambiguous client requirements into working products. Hands-on expertise in RAG, multi-agent orchestration, LLM fine-tuning, full-stack development, and local model deployment.

Professional experience

JMR Infotech Jun 2025 to Present

Lead AI Consultant Bangalore, India

SensAI: Enterprise AI Agent Studio

- Architected and built SensAI end to end, an enterprise platform for building, deploying, and orchestrating AI agents, adopted by 50+ internal users.
- Delivered 6 production agents: LLM Farm, Presales Agent, Knowledge Agent, FSD Generator, Report Summarizer, and Debug Visualizer.
- Developed and optimized RAG pipelines using SentenceTransformers and pgvector, achieving 85%+ query precision with response times of 1 to 3 seconds.
- Implemented multi-agent orchestration with caching, reducing manual effort by over 60%.
- Designed an interactive AI Avatar for client demos and presales that generates virtual presenters, explains project content, and supports live Q&A.
- Built automated test case generation for STDCIF and STDCUSAC and AI-based ALM reporting for OFFSA, maintaining 15-day sprint iterations with zero major rework across a year of delivery.

Ellicium Solutions, an Emergys Company Sep 2023 to May 2025

Data Scientist Pune, India

Wintak/ATAK

- Converted a 3-week client proof of concept into a 7-month production engagement by translating ambiguous requirements into technical specifications and driving ongoing product delivery and new business.
- Migrated the application from Streamlit to Flask, building the backend and frontend from scratch to deliver a fully functional product.
- Built a Text-to-SQL RAG system enabling natural language queries across a schema of 10 to 30 tables, achieving over 70% query accuracy.
- Built an LLM-based report generator connected to the database, producing interactive charts and visuals and reducing analyst report creation time from 2 to 3 days to 20 to 40 minutes.
- Developed drone and vessel threat prediction models achieving an F1 score of 0.71, and a clustering-based algorithm combining RF, IR, and radar sensor data.

Address Matching Tool for CRIF

- Built a vector-based address matching tool using cosine similarity, achieving 90%+ match accuracy across 100k+ addresses. Grouped address clusters by pincode to identify business opportunities and introduced the approach to the client.

KBR: Document question answering

- Built a RAG-based PDF question-answering chatbot with image retrieval. Integrated Llama and Gemma with quantization and LoRA to replace proprietary models, reducing per-query cost from $5–$10 to $0.10–$0.30, with latency of 150 to 400 ms.

Internship experience

Joist AI Feb 2023 to Sep 2023

Data Scientist Intern Remote

- Built an OCR-based pipeline for extracting data from SF330 forms and implemented PDF annotation to support data extraction across 50+ business proposals.

Space Teck Global Sep 2022 to Feb 2023

AI/ML Intern Pune, India

- Built and deployed a Dialogflow chatbot for 20 business clients and LinkedIn sales automation with 10+ filters to streamline prospecting and sales workflows.

Atomic Loops Sep 2021 to Feb 2022

Data Analyst Intern Pune, India

- Built a real-time facial feature extraction algorithm for 68-point detection, developed a QGIS plugin, and published an open-source Python package on PyPI.

Algorithmic Electronics Jul 2020 to Jun 2021

Python Developer Intern Pune, India

- Built corporate email and SMS spam filtering systems and a Twilio-based algorithm for SMS and email communication.

Selected engineering projects

Retail-Banking Ticket Intent Classifier

- Fine-tuned Qwen3-0.6B using LoRA and TRL SFTTrainer on a 26,000-row retail-banking dataset covering approximately 27 intents. Achieved 93.4% accuracy and 100% JSON-schema validity on a golden evaluation set of 2,600 examples, returning structured intent, confidence, and reason fields.
- Built the complete training-to-inference pipeline: dataset preparation, LoRA training on Apple Silicon, adapter merging, GGUF f16 conversion, and local inference through llama-cpp-python. Packaged the solution as an importable Python library.

rusty-agent-sdk

- Built and published a Rust-powered Python SDK for OpenAI-compatible text generation and streaming with zero Python runtime dependencies.
- Implemented blocking and streaming generation, multi-turn conversations, JSON mode, token usage tracking, and provider presets for OpenAI, Anthropic, and OpenRouter.
- Shipped pre-built wheels for Linux, macOS, and Windows with automatic retries, exponential backoff, and configurable timeouts.

Skills

Client delivery: Discovery, technical scoping, requirements translation, stakeholder management, client demos, proof-of-concept development, production delivery.

Applied AI: RAG, Text-to-SQL, multi-agent orchestration, prompt engineering, LLM fine-tuning, LoRA, quantization, local inference, Transformers, HuggingFace, LangChain, LlamaIndex, SentenceTransformers, Ollama, Instructor, pgai/pgvector.

Engineering: Python, FastAPI, Flask, SQLModel, Pydantic, uv, OpenAI SDK, Anthropic SDK, MCP, Git, GitHub.

Data and machine learning: Pandas, NumPy, Scikit-Learn, Keras, TensorFlow, NLTK, Matplotlib, PowerBI, QGIS.

Cloud and databases: AWS, Docker, Linux, PostgreSQL, MySQL, MongoDB, Firebase.

Development tools: Claude Code, OpenCode, OpenAI Codex, GitHub Copilot, Cursor, VSCode.

Education

BE Computer Engineering
Honours in Data Science and Machine Learning
Marathwada Mitra Mandal's College of Engineering 2019 to 2023
CGPA: 8.40/10.00 Pune, India

Certificates

IIRS, ISRO: Remote Sensing and Digital Image Analysis.

AWS Academy Graduate: Data Analytics; Cloud Foundations.

Google Qwiklabs: Cloud Environment; Machine Learning Models in Google Cloud.

Anthropic Academy: Claude 101; Claude Code in Action; Introduction to Claude Cowork.