Akshay Mewada

akshay.mewada7@gmail.com | +91-9821335707

Professional Summary

Senior AI Engineer with 8+ years of experience building production-grade AI, GenAI, recommendation systems, and data-driven platforms across healthcare, finance, and enterprise domains. Specialized in Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), intelligent document processing, and scalable AI pipelines using AWS and Azure. Experienced in architecting conversational AI, hybrid retrieval, OCR, NL-to-SQL, and real-time voice AI solutions focused on scalability, reliability, and business impact.

Experience

First American India

February 2026 - July 2026

Senior Software Engineer II

QCorrect - DES

- Contributed to the redesign of a legacy document processing system into an AI-driven pipeline, improving scalability and performance of document extraction workflows.
- Designed and implemented a multi-model routing approach (custom ML + LLMs + hybrid pipelines) to balance cost, latency, and accuracy based on input complexity.
- Reduced document processing time from ~5 minutes to under 30 seconds while improving extraction accuracy from 70% to 95% through prompt optimization and pipeline improvements.
- Implemented validation and fallback mechanisms to handle low-confidence outputs, improving reliability in production scenarios.

Reliable Group

March 2024 - January 2026

Senior Python Developer - AI Engineer

AskMia - MiaCares

- Led the development of end-to-end RAG-based AI solutions using LLMs for healthcare documents including provider directories, drug formularies, and policy search workflows.
- Architected scalable ingestion pipelines for EOC, CMS formulary, OTC, and provider datasets using Azure Document Intelligence, custom PDF parsers, and distributed processing workflows.
- Improved conversational AI accuracy and contextual understanding through a multi-stage intent orchestration pipeline leveraging fine-tuned DistilBERT, GPT-4o Mini, and Random Forest models.
- Designed hybrid retrieval workflows combining embeddings, lexical search, and adaptive query rephrasing to improve retrieval relevance and reduce hallucinated responses.
- Developed AI-powered provider and pharmacy search solutions using AI agents and NL-to-SQL for natural-language healthcare queries.
- Delivered Mia-Buddy (Image Q&A) and Mia-Choice (Plan Comparison) features leveraging vision models and fallback orchestration workflows.

- Designed and prototyped a real-time VoiceBot using LiveKit and streaming-based architecture for low-latency conversational AI.

Neosoft Technologies *March 2018 - March 2024*

Software Engineer - Associate Team Lead (Data Science)

LatticeFi - Bondlinc

- Led development of AI-powered financial platforms including document summarization, Q&A systems, and intelligent classification workflows for bond analytics, improving user engagement by 12%.
- Built IBondPro, an AI-driven bond portfolio management platform with automated document extraction, semantic search, and insight-generation capabilities, reducing manual research effort for financial documents.
- Developed NLP pipelines for document classification, sentiment analysis, and financial data extraction to support faster trading and investment decision workflows.

ZingHR - AI Module

- Led development of AI-powered recruitment solutions including resume parsing, SkillBank, and candidate-job recommendation systems.
- Built intelligent matching and profile extraction workflows that reduced HR screening effort by ~30% through automated candidate evaluation and personalized recommendations.

HT Media - Slurrp

- Built personalized recommendation workflows for Slurrp using user behavior, cuisine preferences, and contextual signals for recipe and content discovery.
- Developed data processing and ranking pipelines to improve personalized content recommendations and user engagement.

Mahindra and Mahindra - Internal Products

- Integrated AI-powered chatbot solutions across finance, HR, and sales workflows using enterprise communication platforms and WhatsApp integrations via Twilio and Infobip.
- Led computer vision initiatives for tractor decal validation and sealant defect detection, achieving 93.5% accuracy in label detection and 98.5% accuracy in defect-classification workflows.

DeepAIMed - Medical Search Engine

- DeepAIMed is an artificial intelligence medical search engine that brings all healthcare-relevant web content under a unified search experience.
- It retrieves the best results for semantic queries with medically valid concepts and does not look for word matches such as conventional search engines.

Di Jones - Property Recommendation

- Developed an intelligent real-estate recommendation engine that interprets natural-language user queries to deliver personalized property suggestions, integrating location, budget, and lifestyle preferences.

Skills

Programming Languages	Python, JavaScript (Basic)
ML & AI	Machine Learning, Natural Language Processing, Prompt Engineering, Retrieval-Augmented Generation (RAG), Large Language Models (LLMs),
Speech AI:	STT (Whisper, GPT-4o-transcribe), TTS (ElevenLabs)
Web Frameworks	Django, Flask, FastAPI, LiveKit
Databases & Search	SQL, MongoDB, Elasticsearch, Vector Database (Redis, Milvus)
Others	Docker, Kubernetes, Spark, Airflow, Databricks
Cloud	AWS (S3, EC2, Lambda, SageMaker etc.) Azure (Blob Storage, Azure Functions, Azure ML Studio etc.)

Education

K.C. College of Engineering, University of Mumbai *August 2012 - May 2016*

B.E. in Electronics and Telecommunication (6.28/10 CGPA)

Awards and Certification

- Appreciated by Mahindra and Mahindra for playing a crucial role in the Decal and Sealant Project.
- Received recognition for exceptional performance within the data science team during the lockdown
- Certification of Machine Learning by Andrew Ng (Coursera).