

Mohammed Sameer Pasha

+91-93929-74781 | mohammedsameerpasha128@gmail.com | Hyderabad, Telangana, India
[linkedin.com/in/sameer-pashaaa](https://www.linkedin.com/in/sameer-pashaaa) | github.com/sameerpashaaa | sameerpashaaa.github.io/portfolio

Professional Summary

Generative AI Engineer with 1+ year of production experience building LLM-powered applications, Retrieval-Augmented Generation (RAG) systems, and multi-agent AI architectures. Expertise in Python, PyTorch, LangChain, and full-stack development (Django, FastAPI), with proven ability to deploy scalable machine learning models and AI-driven enterprise solutions. Skilled in natural language processing, computer vision, vector databases, and cloud-based model deployment. Delivered AI systems serving 50+ concurrent users and processing 5,000+ documents with sub-3-second response times.

Technical Skills

Programming Languages: Python, SQL, JavaScript, Bash, Git
Generative AI & LLMs: LangChain, Retrieval-Augmented Generation (RAG), Prompt Engineering, Multi-Agent Systems, CrewAI, OpenAI API, Groq, Mistral, Chain-of-Thought Reasoning, Function Calling, LLM Orchestration
Machine Learning: Supervised Learning, Model Training & Evaluation, Feature Engineering, Scikit-Learn, Hyperparameter Tuning, Cross-Validation, Model Optimization, Data Preprocessing
Deep Learning: PyTorch, TensorFlow, Convolutional Neural Networks (CNNs), Transfer Learning, Model Fine-tuning, HuggingFace Transformers, YOLO, Computer Vision
Natural Language Processing: NLP, Sentence Transformers, Embedding Models, Intent Classification, Named Entity Recognition, Text Classification
Speech AI: Speech-to-Text, Text-to-Speech, LiveKit (Voice Agents), Real-Time Audio Processing, Bilingual NLU
Vector Databases: FAISS, ChromaDB, Semantic Search, Embedding Pipelines, Similarity Search
Web Frameworks: Django, Flask, FastAPI, Streamlit, REST APIs, WebSockets, Microservices
Databases: PostgreSQL, MySQL, MongoDB, SQLite, Redis (Caching), Database Optimization, Query Performance Tuning
MLOps & Deployment: Docker, Linux, Git, CI/CD, Model Serving, API Development, Model Versioning, Oracle Cloud Infrastructure, Google Colab, Jupyter Notebook
Data Science & Analytics: Pandas, NumPy, Matplotlib, Seaborn, Data Visualization, ETL Pipelines, Statistical Analysis, Exploratory Data Analysis (EDA)
Computer Vision: OpenCV, Image Processing, Object Detection, OCR (Optical Character Recognition), YOLOv8, EasyOCR, Preprocessing Techniques
Development Tools: VS Code, Postman, GitHub, Kaggle, Jupyter, Git Version Control
Soft Skills: System Architecture Design, Client Communication, Requirement Gathering, Technical Documentation, Agile Development, Cross-Functional Collaboration, Problem Solving

Professional Experience

Metaphasic Life Sciences <i>Software Engineer – AI & Full Stack Development</i>	Hyderabad, India March 2026 – Present
---	--

- Architected and deployed a multi-tenant Enterprise Resource Planning (ERP) system using Django, PostgreSQL, and REST APIs, serving 5 departments with 50+ concurrent users, reducing operational processing time by 40% through automated workflow orchestration and role-based access control.
- Engineered HIPAA-aligned Electronic Health Records (EHR) system using Django and encrypted PostgreSQL, enabling 3 healthcare clinics to digitize 10,000+ patient records with audit logging, appointment scheduling, and prescription management, improving record retrieval time by 90%.
- Delivered real-time inventory management system for 200+ room hotel client using Flask, MySQL, and RESTful APIs, tracking 5,000+ SKUs across 8 departments, reducing stock-outs by 65% and cutting manual reconciliation time from 8 hours to 20 minutes through automated threshold alerts.
- Developed and launched responsive corporate website achieving 95+ PageSpeed score and 300% increase in inbound lead generation within 3 months through SEO optimization and conversion-focused UI/UX design.
- Implemented role-based authentication and authorization system using JWT tokens and Django permissions framework, securing 6 distinct user role hierarchies with granular module-level access control.
- Optimized database query performance by implementing indexed lookups and Redis caching, reducing average API response time from 1.8s to 240ms for high-traffic endpoints serving 1,000+ daily requests.

Projects

StackMind <i>Production RAG System with Multi-Agent LLM Orchestration</i> <i>Technologies: LangChain, RAG, Groq, FAISS, CrewAI, Python, Streamlit</i>	Dec 2024
---	----------

- Engineered production-grade Retrieval-Augmented Generation (RAG) chatbot using LangChain, Groq LLMs, and FAISS vector database, processing 5,000+ technical documents with 89% answer relevance score and sub-3-second response latency for multimodal queries (text, code, diagrams, voice).
- Implemented chain-of-thought prompting with ReAct framework for complex query decomposition, improving answer accuracy by 34% on multi-hop reasoning tasks and enabling transparent step-by-step explanations for technical problem-solving workflows.
- Architected multi-agent AI system using CrewAI framework with specialized agents (Researcher, Planner, Code Analyst, Synthesizer) orchestrating parallel task execution, achieving 40% faster complex query resolution through role-based delegation, shared memory context, and real-time API integration.

- Optimized vector embedding pipeline using Sentence Transformers and semantic chunking strategies, reducing context retrieval time by 60% through sliding-window segmentation with 200-token overlap for long-form technical documentation.
- Built Streamlit web interface with real-time streaming responses, session memory persistence, and multimodal input handling (file upload, voice recording, code snippets), supporting 50+ concurrent user sessions with stateful conversation tracking.

Restaurant AI | *Real-Time Voice Agent with Bilingual NLU & LLM Integration*

Technologies: LiveKit, Mistral API, SQLite, Python, Speech Processing

Nov 2024

- Developed enterprise-ready voice AI agent using LiveKit for real-time speech processing (150ms latency), Mistral LLM for intent classification and dialogue management, supporting English/Hindi bilingual conversations with 92% intent recognition accuracy across 20+ reservation scenarios.
- Engineered intelligent reservation system with Mistral API integration for context-aware dialogue flow, managing a 15-table availability matrix, party-size optimization, and conflict resolution logic, reducing average booking time from 3 minutes (phone) to 45 seconds through natural language automation.
- Implemented full-duplex voice pipeline with LiveKit's WebRTC streaming (speech-to-text + text-to-speech), achieving 150ms round-trip latency and 94% transcription accuracy (English/Hindi), with intelligent seating recommendation engine integrating OpenWeatherMap API for indoor/outdoor preference optimization.
- Designed SQLite database schema with normalized tables for reservations, customer history, and table status, implementing transaction-safe CRUD operations and concurrent booking conflict detection with row-level locking.

ANPR System | *End-to-End License Plate Detection with YOLOv8 & OCR Pipeline*

Technologies: YOLOv8, OpenCV, EasyOCR, PyTorch, Python, Computer Vision

May 2024

- Engineered end-to-end Automatic Number Plate Recognition (ANPR) pipeline using YOLOv8 for real-time detection and EasyOCR for text extraction, trained on 15,000+ annotated images (Kaggle dataset), achieving 91% plate localization mAP@0.5 and 87% character recognition accuracy at 18 FPS on CPU.
- Developed adaptive preprocessing pipeline with Gaussian blur, Otsu's binarization, morphological transformations, and CLAHE enhancement for low-light/skewed plates, boosting OCR accuracy from 62% to 87% (+25 percentage points) and reducing false positives by 40% through contour area filtering.
- Optimized inference pipeline with frame skipping (process every 3rd frame), batched preprocessing, and ROI-focused detection zones, reducing per-frame processing time from 5.2s to 1.8s (65% improvement) and achieving 18 FPS throughput for real-time video stream analysis.
- Implemented real-time visualization dashboard using OpenCV and Streamlit, displaying bounding boxes, OCR confidence scores, and recognized plate numbers with 50-frame history tracking and CSV export functionality for analytics and validation.

Education

Muffakham Jah College of Engineering and Technology	Hyderabad, India
<i>B.Tech in Artificial Intelligence and Machine Learning</i>	<i>Aug 2022 – Aug 2026</i>
Sri Chaitanya Junior College	Hyderabad, India
<i>Intermediate (MPC) – Mathematics, Physics, Chemistry</i>	<i>June 2020 – May 2022</i>

Certifications

Oracle Cloud Infrastructure 2024 AI Foundations Associate	Sept 2024
<i>Certified in OCI AI fundamentals, cloud AI services, deployment workflows, and responsible AI frameworks</i>	
AWS Solutions Architecture Virtual Experience	July 2024
<i>Designed scalable AWS Elastic Beanstalk architecture with auto-scaling, load balancing, and cloud best practices</i>	
Deloitte Technology Job Simulation	June 2024
<i>Built performance monitoring dashboard and delivered technical recommendations to stakeholders</i>	
100 Days of Code: Python Pro Bootcamp (Udemy)	May 2024
<i>Comprehensive Python development course covering data structures, OOP, web frameworks, and automation</i>	