

RAVURI SAI VIDYASAGAR

AI Engineer · Generative AI Engineer · LLM Solutions Architect

 sagarravuri2@gmail.com |  +91-7993425279 |  Hyderabad, Telangana |  [linkedin.com/in/sagar-ravuri-b048151a0](https://www.linkedin.com/in/sagar-ravuri-b048151a0)

PROFESSIONAL SUMMARY

Generative AI Engineer with 4+ years of hands-on experience designing and deploying production-grade LLM systems, RAG architectures, and agentic AI pipelines in healthcare. Deep expertise in LLM fine-tuning, prompt engineering, vector databases, OCR-driven document intelligence, and multi-modal AI workflows. Proven track record of reducing operational cycle times by up to 40% and scaling GenAI systems to 1,000+ daily transactions at 99.5% uptime. Passionate about building intelligent, explainable, and scalable Generative AI solutions that solve real-world problems.

KEY ACHIEVEMENTS

-  40% reduction in healthcare workflow processing time through LLM-based automation
-  30% improvement in document processing accuracy via integrated OCR + LLM pipelines
-  Scaled ML systems to 1,000+ daily patient case submissions at 99.5% uptime
-  5,000+ daily active users served across production AI systems in healthcare organizations
-  70% reduction in ML development time by building a no-code AI platform (CIBI) for 200+ users

TECHNICAL SKILLS

LLMs	GPT-4, LLaMA 3, Qwen, BERT · Fine-Tuning · RLHF · Prompt Engineering · Text Classification
GenAI & RAG	Retrieval-Augmented Generation · LangChain · Agentic AI · Multi-turn Dialogue · Chain-of-Thought Prompting
Vector Databases	Pinecone · FAISS · Semantic Search · Embeddings · Similarity Retrieval
OCR & Doc AI	LayoutLM · OpenCV · Vision-Language Models · Multi-modal Document Intelligence
AI Infra & APIs	FastAPI · Flask · REST APIs · Microservices · Docker · MongoDB · Python (advanced) · Git
GenAI Frameworks	HuggingFace Transformers · PyTorch · Scikit-learn
Healthcare GenAI	Prior Authorization Automation · Clinical NLP · ICD-10/CPT Extraction · EHR Integration · HIPAA

PROFESSIONAL EXPERIENCE

Generative AI Engineer | PenguinAI India — Hyderabad

July 2024 – Present

- Architected end-to-end AI workflows combining LLMs, RAG, OCR, and agent-based reasoning for healthcare payers and providers
- Reduced prior authorization and claim review handling times by 30–40% through automated, scalable pipeline design
- Deployed multi-modal data ingestion pipelines covering EHR data, scanned PDFs, medical forms, and clinical notes
- Defined BI roadmap and technology vision aligned with corporate objectives; collaborated with senior management on risk assessment and audit planning
- Delivered AI-driven recommendation engines and store viability models, communicating findings clearly to stakeholders

Key Projects:

▶ HCS Healthcare – Prior Authorization Automation

- Designed Qwen-based dynamic questionnaire engine with Vector DB RAG for real-time guideline retrieval
- Achieved 40% reduction in manual PA processing with 25% faster approval turnaround; 95%+ retrieval accuracy
- Scaled to 500+ authorization requests daily across multiple payer guidelines via Pinecone vector database

▶ Respire – Nutritional Prescription Processing

- Engineered OCR + NLP workflow to extract demographics, diagnosis codes, and nutritional plans from prescriptions with 98% accuracy
- Saved 12+ hours/week in manual transcription, improving operational efficiency by 85%

▶ Snowflake PA – Prior Authorization Decision Agent

- Built agentic pipeline combining OCR, ICD-10/CPT mapping, and eligibility validation with transparent audit UI
- Deployed FastAPI microservice with MongoDB integration handling 300+ concurrent API requests at sub-second response time
- Accelerated authorization decisions by 30% while maintaining full regulatory compliance

▶ Ovation Billing Review & DPP

- Engineered digital billing agent using RAG architecture for automated CPT/ICD extraction and structured YAML output
- Built and deployed XGBoost claim denial prediction model (87% precision) via FastAPI with SHAP explainability
- Applied advanced feature engineering on 50,000+ historical claims for transparent, production-ready model interpretability

▶ Respire DPP – Patient Risk Scoring

- Designed risk stratification model to categorize patients into nutritional risk tiers, generating personalized dietary recommendations

AI Engineer | Awone AI — Hyderabad

March 2023 – June 2024

- Built comprehensive Python ML toolkit via FastAPI microservices enabling users to develop production-grade ML/DL models with minimal code
- Spearheaded synthetic data generation and architected monitoring dashboards with optimized MongoDB schemas
- Empowered 200+ users, reducing ML development time by 70% through automated preprocessing pipelines
- Enforced data modelling standards and best practices; contributed to design reviews and system architecture
- ▶ CIBI – No-Code AI Platform
 - Designed end-to-end no-code interface for ML/DL model training, evaluation, and deployment accessible to non-engineers
- ▶ Maps – Government Maps Digitization
 - Created OCR framework for extracting numeric depth values from bathymetric maps with $\pm 50\text{m}$ accuracy
 - Reduced manual digitization time by 50% for marine authority workflows

AI & Software Engineer | Pyro Holdings — Hyderabad

August 2021 – February 2023

- Orchestrated end-to-end data pipelines and automated reporting using Python and SQL, improving turnaround by 35%

EDUCATION

Bachelor of Engineering – Computer Science & Engineering

Sathyabama Institute of Science and Technology, Chennai | 2017 – 2021

CORE COMPETENCIES

LLM Architecture & Fine-Tuning · RAG System Design · Agentic AI Pipelines · Prompt Engineering · OCR & Document Intelligence · Vector Databases (Pinecone, FAISS, Weaviate) · LangChain · GenAI for Healthcare (HIPAA) · FastAPI Microservices · Multi-modal AI · Scalable AI Pipeline Design