

Rishav Mallick

Senior AI/ML Engineer | Clinical & Healthcare AI Systems

rishav@rmlk.dev | rmlk.dev | linkedin.com/in/rishavmallick

Summary

AI/ML engineer with 9 years of experience in NLP, speech, and computer vision, working mainly on healthcare systems in recent years. I build and ship production inference services (medical RAG, speech recognition, structured extraction) with safety checks, audit logging, and observability included from the start.

Skills

Backend: Python, FastAPI, Flask, SQL, MySQL, Nginx, Git, JavaScript

LLM & retrieval: vLLM, RAG, FAISS, BM25 hybrid search, cross-encoder reranking, JSON-schema structured output, sentence-transformer fine-tuning

Speech & vision: Whisper ASR (GPU inference, model pooling, VAD), XTTS-v2 TTS, Indic-language ASR/TTS, OCR, OpenCV

ML: PyTorch, TensorFlow, scikit-learn, XGBoost, Hugging Face Transformers, ClinicalBERT

Cloud & MLOps: AWS (SageMaker, S3, EventBridge), MLOps, structured logging, production-readiness audits

Experience

Independent AI/ML Engineer | Freelance Present

- Built FastAPI inference services for clinical use: a GPU-backed Whisper ASR service with model pooling, denoising, and VAD, and an XTTS-v2 speech synthesis server for English, Hindi, and Punjabi.
- Designed a medical RAG patient-safety chatbot backend (intent classification, safety policy, FAISS + cross-encoder retrieval, GPT-4.1, structured audit logging) and ran a production-readiness audit that produced a 21-step remediation plan.
- Built a four-pass concurrent vLLM pipeline that turns consultation transcripts into about 18 categories of structured EMR data; validated on Llama 3.1 8B and MedGemma-4B, targeting 100 concurrent users on a single L40S GPU.
- Built a hybrid FAISS + BM25 drug search service with hospital-aware prescription enrichment, and a token-level anomaly detector for clinical transcripts (ClinicalBERT outperformed GatorTron and Bio_ClinicalBERT in benchmarking).

Senior Data Engineer | Global IT services firm, India Nov 2024 – Sep 2025

- Built and validated Logistic Regression and Random Forest models for early detection of pre-eclampsia, working with a US healthcare client's clinical team; extracted and prepared data with Python, Pandas, and PySpark.
- Designed a predictive health-risk alert pipeline (XGBoost, Random Forest, CTGAN synthetic data for class imbalance) with real-time and batch inference, using SageMaker, S3, and EventBridge.
- Worked under clinical data governance requirements (secure handling, auditability, versioning); presented progress in daily client meetings and delivered project documentation.

Senior Project Engineer | Government R&D organization, India Apr 2017 – Oct 2024

- Led development of a Bengali and Hindi multimodal question-answering chatbot: fine-tuned sentence transformers (Siamese networks), built a Flask data-collection app, and customized ASR and TTS for Indic languages.
- Built an Odia document OCR system (OpenCV preprocessing, CNN-LSTM model, image-to-text API and GUI) reaching 94% character-level accuracy.
- Led data acquisition from special schools, NGOs, and regular schools for an autism screening app for children aged 2 to 12; Random Forest reached 97% accuracy against Naive Bayes, KNN, and SVM.

Education

PG Diploma, Big Data Analysis 2017

B.Tech, Computer Science Engineering 2015