

AMAN JAISWAL

Bengaluru, India | Mob:+91-8084410670 | ai.amanjaiswal@gmail.com | www.linkedin.com/in/amanjaiswal777/

EXPERIENCE

ClarismD - Self-Funded, Solo-Built - India

Founder & Principal AI Architect (Sole Engineer)

Nov 2025 – Present

Live: clarismd.com · SDKs: [@clarismd/sdk \(npm\)](https://www.npmjs.com/package/@clarismd/sdk) · [clarismd \(PyPI\)](https://pypi.org/project/clarismd/)

Key Achievements

- Built and launched ClarismD, a production healthcare AI governance platform that secures and governs interactions with OpenAI, Anthropic, Gemini, and AWS Bedrock through a unified LLM gateway.
- Architected and operate an end-to-end AI governance platform, including a FastAPI gateway, governance dashboard, and public SDKs for Python and TypeScript.
- Built a multi-provider LLM gateway supporting OpenAI, Anthropic, Gemini, and AWS Bedrock with tenant isolation, encrypted key management, semantic caching, rate limiting, cost attribution, and budget controls.
- Developed healthcare-grade AI safety and compliance capabilities including PHI/PII detection and redaction, prompt injection defense, toxicity validation, constitutional AI evaluation, and automated red teaming.
- Implemented governance and compliance controls across 17+ frameworks including HIPAA, HITECH, GDPR, EU AI Act, NIST AI RMF, ISO 42001, SOC 2, FDA SaMD AI/ML, and ONC HTI.
- Designed governance modules for RAG and AI agents, including hallucination detection, retrieval monitoring, agent-loop detection, memory governance, and audit-ready evidence generation.
- Established enterprise-grade security, observability, and compliance infrastructure with RBAC, MFA, audit trails, OpenTelemetry, Prometheus, Grafana, Sentry, and automated CI/CD pipelines.

Tools/Technologies Used:

- FastAPI, Python, React, TypeScript, PostgreSQL, Redis, Qdrant, OpenAI, Anthropic Claude, Gemini, AWS Bedrock, Docker, GitHub Actions, OpenTelemetry, Prometheus, Grafana, Sentry.

Rigetnest Innovation Labs, India

Co-Founder & Principal AI Architect

Aug 2023 – Present

Website: rigetnest.com

Key Achievements:

- Co-founded an AI and product engineering company focused on enterprise software, Generative AI, and digital transformation solutions.
- Led stakeholder discovery, solution architecture, technical roadmapping, effort estimation, and delivery execution across multiple client engagements spanning AI, SaaS, and enterprise applications.
- Architected and guided the development of ResumeX, an AI-powered resume intelligence platform enabling automated resume analysis, candidate profiling, and career optimization workflows.
- Designed and delivered Orange Technology's business platform, translating complex business requirements into scalable product architectures, implementation roadmaps, and engineering deliverables.
- Established engineering standards and reusable AI frameworks while mentoring teams and driving successful delivery of cloud-native, AI-powered, and enterprise-scale solutions.

Technologies:

- Generative AI, LLMs, Python, FastAPI, React, TypeScript, AWS, Docker, PostgreSQL, System Design, Solution Architecture, Product Strategy, Technical Consulting.

United Airlines (KForce), USA (India Remote)

Senior LLM Scientist (Gen AI Consultant)

May 2025 – Present

Key Achievements:

- Architected an AI-driven Weekly Status Report (WSR) automation platform integrating Jira portfolio analytics, S3-versioned PowerPoint templates, and a proxy-routed LLM layer to generate executive-ready PPTX reports with consistent, audit-friendly metrics.
- Engineered resilient LLM and reporting pipelines (fallbacks, deterministic text fitting, metric reconciliation/observability, multi-portfolio & business-unit rollups), demonstrating expertise in LLM systems design, data governance, and production automation.
- Built an enterprise AI customer service response system for United Airlines that generates context-aware smart responses using dual-summary architecture (brief for LLM context, detailed for human output) with multi-layered safety guardrails. Implemented production-ready infrastructure with parallel multi-model inference, real-time PII masking, prompt injection detection, toxicity validation, and strict output format enforcement for secure, compliant AI responses at scale.
- Architected a production LLM-driven ground-services invoice reconciliation platform for United Airlines line stations, scaling station-specific builders across cabin cleaning, VBC, wheelchair, ground handling, and hotel/meal-voucher domains by integrating a Gemini PDF→Markdown + Bedrock Claude Sonnet 4.6 schema-extraction pipeline with a LangGraph multi-agent workflow and in-tree LightRAG knowledge graph (Neptune + Milvus + Cohere Rerank-3.5) over a tri-database backend (Postgres MARS RDS, NetOps Redshift, Deep Clean Redshift), enabling finance to automatically validate vendor PDF invoices against UA-computed service classifications at airport-by-airport scale.

Tools/Technologies Used:

- FastAPI, AsyncIO, Poetry, Jinja2, LangChain, OpenAI GPT, Anthropic Claude Sonnet 4.5, AWS Bedrock, Kubernetes, Helm, Docker, AWS ECR, AWS S3, Arthur Shield API, PII Detection, Prompt Injection Detection, Toxicity Validation, Keyword Filtering, OAuth2, Bearer Token Authentication, Multi-model Ensemble Inference, Async Batch Processing, Dual-Summary Architecture, Request Tracing, Token Usage Tracking, Error Handling, Middleware Authentication

KovrAI (Supertal.io), Gurugram, India

Senior Data Scientist (Gen AI Consultant)

May 2025 – Oct 2025

Key Achievements:

- Designed and led the implementation of an enterprise-grade Retrieval-Augmented Generation (RAG) platform for AI compliance and policy automation, enabling context-aware document retrieval, control mapping, and LLM-driven recommendations across regulated frameworks such as FedRAMP and CMMC.
- Incorporating NIST 800-53, CMMC, and other frameworks into LLM prompting logic.

Tools/Technologies Used:

- Python, NLP, Deep Learning, Transformers, FASTAPI, Streamlight, Docker, Kubernetes, LLMs, MilvusDB Sentence-transformers, Chainlit, LLama, OpenAI, AWS, Azure, GCP, Langchain, VLLMs, MLFlow, MLOps

Xenie-AI (Freelance Engagement), India (Remote)

Freelance AI Engineer – Voice & Multimodal AI

Apr 2025 – Oct 2025

Key Achievements

- Architected and delivered a multilingual voice AI assistant supporting 9 Indian languages, leveraging NVIDIA NeMo, IndicF5, and end-to-end STT → LLM → TTS pipelines for real-time conversational experiences.
- Built a multimodal emotion intelligence platform combining text, voice, and facial signals using BERT, Wav2Vec2, DeepFace, and ensemble learning for unified emotion analysis.
- Developed voice biometrics and wake-word detection systems using ECAPA-TDNN, Wav2Vec2, CNN-LSTM, and real-time audio processing for secure, hands-free user interactions.
- Designed and deployed a RAG-powered conversational assistant with cross-lingual retrieval, semantic search, and language-aware knowledge access using LangChain, ChromaDB, and transformer-based embeddings.
- Engineered scalable microservices and modern web applications using FastAPI, WebSockets, React/Next.js, and AWS infrastructure to support low-latency, production-grade voice AI workloads.

Tools/Technologies Used

- Python, FastAPI, NVIDIA NeMo, IndicF5, PyTorch, TensorFlow, Wav2Vec2, BERT, ECAPA-TDNN, DeepFace,

LangChain, ChromaDB, OpenAI, Next.js, React, TypeScript, AWS, Docker.

EONMED (Freelance Engagement), India (Remote)

Freelance Consultant (Part-time) - Healthcare AI

Feb 2025 – Jun 2025

Key Achievements:

- Architected and delivered an AI-powered healthcare platform for non-invasive vital sign monitoring, enabling estimation of heart rate, blood pressure, SpO₂, respiratory rate, and stress levels directly from smartphone video.
- Developed computer vision and rPPG-based pipelines for physiological signal extraction, combining signal processing, machine learning, and demographic modeling for remote health assessment.
- Built multi-modal medical AI solutions for cataract grading, skin cancer classification, eye condition detection, tongue diagnosis, and stress analysis using CNNs, Vision Transformers, YOLOv8, and explainable AI techniques.
- Integrated medical LLM capabilities through FastAPI services, enabling specialty-specific image interpretation workflows across radiology, dermatology, pathology, and ophthalmology use cases.
- Engineered scalable healthcare AI microservices with asynchronous processing, optimized inference pipelines, AWS-backed data workflows, and production-ready APIs for mobile and web platform integration.
- Designed AI-driven healthcare and fintech solution concepts, including multilingual voice-based assistants, medical reasoning workflows, and speech-to-query systems for real-world deployment scenarios.

Tools/Technologies Used:

- Python, FastAPI, PyTorch, TensorFlow, OpenCV, MediaPipe, YOLOv8, Vision Transformers, MedGemma, Hugging Face, AWS S3, Docker, Scikit-learn, XGBoost.

4CRisk.ai, Bangalore, India

Senior Data Scientist (Gen AI)

Jun 2023 – May 2025

Data Scientist

Nov 2022 – May 2023

Key Achievements:

- Developed an AI-powered Platform-as-a-Service (PaaS) that streamlines daily operations for backend and frontend teams, enhancing efficiency and productivity. Additionally, I supported Product Managers (PMs) in conducting Proof of Concepts (PoCs) to explore and validate new ideas, fostering innovation and collaboration across all teams.
- Led the design and development oversight of an AI-driven conversational system, ARIA Co-pilot (RAG - Retrieval-Augmented Generation), in the finance sector, working closely with a small team to enhance functionality and performance.
- Spearheaded a team to scale an ARIA Co-pilot (RAG) application powered by Large Language Models (LLMs) to support over 10k+ users, optimizing performance, reliability, and responsiveness through efficient infrastructure scaling and algorithmic improvements.
- Enhanced data parsing, query extraction, retrieval, ranking, and generation to improve system accuracy, achieving over 70% accuracy as verified by the client.
- Developed and deployed open-source LLMs for in-house consumption along with LLM fine-tuning.
- Designed and supervised the development of an ML inference pipeline, increasing speed by 200% and enabling scalable load handling.
- Established data versioning and experiment tracking pipelines using DVC and MLflow.
- Developed an automated system using semantic analysis to streamline policy alignment, enhance efficiency, and ensure regulatory compliance.
- Trained a deep learning model based on sentence transformers, achieving a remarkable 70% improvement in accuracy.
- Created specialized classification and synthetic data generation APIs, reducing task time by 50%.
- Conducted T5-based model training for extractive summarization with high ROUGE-N scores.
- Engineered a semantic search engine with advanced clustering techniques for improved accuracy.
- Enhanced compliance map accuracy from an F1 score of 30 to over 70 through intensive research.
- Crafted stakeholder experimentation and visualization APIs for seamless deployment.
- Established end-to-end data creation pipelines using LLMs to address data shortages.
- Played a pivotal role in in-house LLM development for text generation and similarity tasks.
- Functioned as a prompt engineer, optimizing prompts for project requirements.

- Led deployment efforts including API creation and Kubernetes deployment, optimizing for cloud providers.
- Initiated and led a pivotal project on classification tasks, leveraging diverse language models for binary and multi-class classification.

Tools/Technologies Used:

- Python, NLP, Deep Learning, Transformers, FASTAPI, Streamlight, Docker, Kubernetes, LLMs, Sentence-transformers, Chainlit, LLama, OpenAI, AWS, Azure, GCP, Langchain, VLLMs, MLFlow, MLOps, MilvusDB

Acidaes Solutions Pvt. Ltd. (CRMNext) - Noida, India

Machine Learning Engineer (MLE)

April 2021 – Oct 2022

Key Achievements:

- Worked on NLP tasks, including text summarization, NER (BERT, Flair, Spacy), sentiment analysis (BERT & MURIL), text classification, and text generation for banking use cases.
- Utilized Pycaret AutoML Framework for Classification, Regression, and Clustering tasks.
- Conducted Emotion Recognition via Speech and implemented real-time facial matching using ML, DL, OpenCV, Tkinter, etc.
- Developed an IDM (Industrial Data Model) Service for banking use cases using MongoDB, Python, and FastAPI.
- Contributed to generating and consuming dummy data for banking customers through a database.
- Managed bulk data through MongoDB, Cassandra, and Pyspark, and worked on Monolingual and Multilingual Automatic Speech Recognition for Indic Languages.
- Deployed services via RestfulAPI and gRPC, created an NLP Chatbot service using RASA, and performed a comparative analysis of data retrieval latency using PySpark.

Tools/Technologies Used:

- Python, NLP, ML, DL, OpenCV, Tkinter, MongoDB, FastAPI, Pycaret AutoML, Cassandra, Pyspark, RASA, PySpark

US Tech Solutions Pvt. Ltd. – Noida, India

Machine Learning Engineer (MLE)

Jan 2020 – March 2021

Key Achievements:

- Developed an innovative resume parser integrating NLP, Computer Vision, and Deep Learning.
- Devised a unique method for extracting sequential text from all formats, enhancing categorization.
- Engineered NER Models for precise data extraction and segmentation.
- Optimized parsing algorithm by 50%, reducing the average time to 2.3s.
- Achieved 90% accuracy in text segmentation and 95% accuracy in data extraction.
- Identified entity relationships in resume segments, improving parsing accuracy.
- Conducted comparative analysis of text extraction libraries, enhancing performance.

EDUCATION

KIET Group of Institutions – Ghaziabad, Uttar Pradesh, India

2016 - 2020

Bachelor of Technology, Computer Science; Cumulative Score: 8.79

Sunbeam School – Varanasi, Uttar Pradesh, India

2013 - 2015

Intermediate, CBSE, PCME Score: 94.25%

SKILLS AND CERTIFICATIONS

Technical: Python, Data Science, Machine Learning, Computer Vision, Deep Learning, NLP, Docker, Rancher, FastAPI, MongoDB Database, gRPC, Signal Processing, Design Patterns, Data Structures, and Algorithm

Tools/OS: Linux, Git, Jupyter-notebook, Spyder, VS code, DL-frameworks (Tensorflow, Keras, PyTorch), Pycharm, AWS Sagemaker, Google Colab, MongoDB Compass, PostMan, Azure, DataGrip

Packages: Numpy, Pandas, SciPy, NLTK, Matplotlib, Seaborn, OpenCV, Tensorflow, Keras, Pytorch, Transformers, Librosa, Spacy, Huggingface

Certifications & Training: Computer Vision Nanodegree (Udacity), Deep Learning Part -1 (NPTEL) and Deep Learning Part -2 (NPTEL), Machine Learning (Coursera), Practical Machine Learning using Tensorflow 2.0 (NPTEL), Introduction to Machine Learning in Production (DeepLearning.ai), Programming Data Structure and Algorithms (NPTEL), Python for data science and machine learning (Udemy), Introduction to programming in C (NPTEL), LLMOps: Building Real-World Applications With Large Language Models (Udacity)

RELEVANT PROJECTS

Crop Disease Detection, Eng2SQL generation(Freelance)

2021

- Worked on detecting diseases in crop leaves and providing additional support to help solve the problem.
- I worked on a CNN-based architectural model for detection that was trained on AWS Sagemaker.
- I worked on customer service needs to generate SQL queries from English text using OpenAI and FastAPI.
- I also worked on Intent Classification and Paraphrasing of English sentences using OpenAI.

Facial Keypoint Detection, Image Captioning & Landmark Detection, Computer Vision Nanodegree, Udacity

May 2020

- Developed a facial keypoint detection system incorporating a face detector leveraging Haar Cascades and a Convolutional Neural Network to predict facial key points in an observed face (CNN).
- Implemented SLAM to track a robot's position in real time and detect landmarks such as buildings, trees, rocks, and other objects in a two-dimensional world.
- We are working on creating and training the CNN-RNN model for automatic picture caption creation in the second project (Convolutional Neural Network - Recurrent Neural Network).

Speaker Verification and Validation (KIET)

Dec 2018 - Feb 2019

- In a team of three, I concentrated on building the general flow and implementing a text-independent speaker recognition system using CNN architecture to distinguish sounds and identify comparable audio spectrograms. MFCCs, CNN, HMM, and GMM models were used for feature extraction, while SVM was used to classify the audio input.

Next-Generation Traffic Light (KIET)

Aug 2018 - Nov 2018

- The project aimed to reduce traffic congestion by allocating dynamic time on each lane based on vehicle density rather than static time.
- I worked on the end-to-end process in a two-person group. Using OpenCV, YOLO, and R-CNN, handled the mathematical analysis of how Object Recognition works for a wide range of tasks, including object position, detection, and image classification.

RESEARCH PUBLICATION & HONOURS

- Artificial Intelligence Analysis of Cultural Narratives Shaping Emotional Responses to Infertility. (<https://ieeexplore.ieee.org/abstract/document/11167416>)
- Analysis of Deep Learning algorithms on COVID-19 Radiography Database (IJAST) in which we proposed a method to detect the COVID-19 presence based on the results of the DL architectures. (<https://github.com/amanjaiswal777/Analysis-of-Deep-Learning-algorithms-on-COVID-19-Radiography-Databases>)
- AUTHEER : A Voice-Based Speaker Authentication System (IJAST) based on speaker identification using CNN. (<https://github.com/amanjaiswal777/AUTHEER-A-Voice-Based-Speaker-Authentication-Systems>)
- **ICCIT 2025 - 4th International Conference on Creative Communication and Innovative Technology** - The conference brought together international researchers, academics, and industry professionals, providing a platform to share cutting-edge developments in AI technology and innovative communication solutions with the next generation of technologists.