

# AMAN JAISWAL

## GenAI / LLM Systems Architect | Independent Consultant

Bengaluru, India (works US & EU hours) | +91 80844 10670 | [ai.amanjaiswal@gmail.com](mailto:ai.amanjaiswal@gmail.com)  
[linkedin.com/in/amanjaiswal777](https://linkedin.com/in/amanjaiswal777) | [github.com/amanjaiswal777](https://github.com/amanjaiswal777)

### PROFESSIONAL SUMMARY

GenAI / LLM Systems Architect with 6+ years in applied AI/ML, including 4 years shipping production generative-AI systems. I help enterprises move LLM prototypes into reliable, secure production: document intelligence, retrieval-augmented generation (RAG), and multi-model AI services built for scale, auditability, and graceful failure. Currently architecting a GenAI document-intelligence and invoice-reconciliation platform for a Fortune 500 airline; previously delivered RAG platforms for financial-services risk and US federal compliance (FedRAMP, CMMC, NIST 800-53). Available for remote consulting engagements, 20-40 hours per week, with daily US Eastern overlap.

### CONSULTING FOCUS

- **LLM & RAG Systems:** production retrieval architecture, document intelligence, structured extraction, and AI microservices taken from proof of concept to production.
- **Reliability & Guardrails:** multi-model fallback and failover, LLM evaluation, real-time PII masking, prompt-injection and toxicity defense, strict output-format enforcement.
- **Enterprise AI Architecture:** enterprise SSO (OAuth2 / OIDC), role-based access control, audit logging, compliance-grade workflows, and Kubernetes delivery.
- **Cost & Performance Optimization:** model routing, token budgeting, multi-model strategy, and inference throughput tuning.

### PROFESSIONAL EXPERIENCE

#### Senior LLM Scientist (GenAI Consultant)

Oct 2025 - Present

*United Airlines (via KForce), USA - Remote from India*

- **Invoice Intelligence & Reconciliation Platform:** architected and own end-to-end a GenAI platform that automates vendor billing validation across multiple airport stations and 8 ground-handling service lines, replacing manual invoice review with automated extraction, reconciliation, and multi-stage approval.
- **Resilient LLM extraction pipeline:** built a PDF-to-structured-data pipeline on Google Gemini behind an enterprise API gateway, with a three-model fallback cascade, page-chunked processing with sliding-window context, token budgeting, and explicit recovery from content-filter and truncation failures, converting brittle single-shot parsing into a production-reliable path.
- **Financial reconciliation & savings engine:** engineered period-aware savings accounting (potential versus confirmed), cross-period rebooking, multi-currency normalization, discrepancy and overrun detection, and content-hash duplicate control, reconciling invoices against operational truth in a cloud data warehouse.
- **Enterprise authentication, RBAC & audit:** implemented enterprise SSO (OIDC and JWT with JWKS validation) behind an API gateway, hierarchical scope-based role-based access control across local, regional, and global scopes, plus actor attribution and lifecycle audit logging to satisfy internal audit requirements.
- **Executive reporting automation:** architected an AI-driven Weekly Status Report platform integrating portfolio analytics, versioned presentation templates, and a proxy-routed LLM layer to produce executive-ready decks with audit-friendly metrics, deterministic text fitting, metric reconciliation, and multi-portfolio rollups.
- **AI customer-service response system:** built a context-aware response system on a dual-summary architecture (compact for model context, detailed for human review) with layered guardrails: real-time PII masking, prompt-injection detection, toxicity validation, and strict output-format enforcement, served via parallel multi-model inference.

**Technologies:** Python, FastAPI, SQLAlchemy, React 18, Google Gemini, Claude Sonnet 4.5, AWS Bedrock, LangChain, PostgreSQL, Amazon Redshift, Arthur Shield, Kong API Gateway, Microsoft Entra ID, Kubernetes, Helm, GitHub Actions

## Senior Data Scientist (GenAI Consultant)

May 2025 - Oct 2025

*KovrAI (Supertal.io), Gurugram, India*

- Designed and led an enterprise RAG platform for AI compliance and policy automation, enabling context-aware document retrieval, security-control mapping, and LLM-driven remediation recommendations across FedRAMP and CMMC.
- Encoded NIST 800-53 and CMMC control logic directly into prompting and retrieval strategy, making model outputs traceable to specific regulatory clauses.

## Senior Data Scientist, GenAI (promoted from Data Scientist)

Nov 2022 - May 2025

*4CRisk.ai, Bangalore, India*

- Led design and delivery of ARIA Co-pilot, a RAG conversational system for financial-services risk, and scaled it to 10,000+ users through infrastructure and algorithmic optimization.
- Rebuilt the retrieval stack (parsing, query extraction, ranking, and generation), reaching 70%+ end-to-end accuracy as independently verified by the client.
- Raised regulatory compliance-mapping quality from an F1 score of 30 to over 70 through targeted research on semantic alignment.
- Designed an ML inference pipeline delivering a 200% throughput increase under scaled load.
- Deployed and fine-tuned open-source LLMs for in-house text generation and similarity tasks, reducing external API dependency and per-request cost.
- Built classification and synthetic-data-generation APIs that cut task turnaround 50%, and established DVC and MLflow pipelines for data versioning and experiment tracking.

## Machine Learning Engineer

Apr 2021 - Oct 2022

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

- Shipped NLP for banking: summarization, named-entity recognition (BERT, Flair, spaCy), multilingual sentiment analysis (BERT, MuRIL), classification, and text generation.
- Built an Industrial Data Model service (MongoDB, Python, FastAPI), deployed ML services over REST and gRPC including a RASA conversational service, and engineered large-scale pipelines across MongoDB, Cassandra, and PySpark with multilingual ASR for Indic languages.

## Machine Learning Engineer

Jan 2020 - Mar 2021

*US Tech Solutions Pvt. Ltd., Noida, India*

- Built a production resume-parsing engine combining NLP, computer vision, and deep learning, reaching 90% text-segmentation and 95% data-extraction accuracy with custom NER models.
- Optimized the parsing algorithm by 50%, cutting average latency to 2.3 seconds.

## TECHNICAL SKILLS

**GenAI & LLM Systems:** LLM application architecture, Retrieval-Augmented Generation (RAG), hybrid retrieval, re-ranking, chunking strategy, document intelligence, structured extraction, multi-model routing and fallback, LLM guardrails, LLM evaluation, token and cost optimization, fine-tuning, prompt architecture, agentic workflow design

**Models & Frameworks:** Anthropic Claude, OpenAI GPT, AWS Bedrock, Llama, vLLM, Hugging Face Transformers, LangChain, Sentence-Transformers, PyTorch, TensorFlow, Google Gemini

**Languages & Backend:** Python, JavaScript, SQL, FastAPI, AsyncIO, Pydantic, SQLAlchemy, REST, gRPC

**Data & Storage:** PostgreSQL, Amazon Redshift, MongoDB, Cassandra, Milvus, OpenSearch, PySpark

**Cloud & MLOps:** AWS (Bedrock, S3, ECR, Redshift, Secrets Manager, CloudFormation, SageMaker), Azure, GCP, Kubernetes, Helm, Docker, GitHub Actions, MLflow, DVC, pytest, mypy, ruff, Linux, Git

**Security & Compliance:** OAuth2, OIDC, JWT and JWKS validation, Microsoft Entra ID, API gateway patterns (Kong), hierarchical RBAC design, audit logging, FedRAMP, CMMC, NIST 800-53

## EDUCATION & CERTIFICATIONS

---

### **Bachelor of Technology, Computer Science**

2016 - 2020

*KIET Group of Institutions, Ghaziabad, India - CGPA 8.79 / 10*

**Certifications:** LLMOps: Building Real-World Applications With Large Language Models (Udacity); Introduction to Machine Learning in Production (DeepLearning.AI); Computer Vision Nanodegree (Udacity)

## PUBLICATIONS

---

- Artificial Intelligence Analysis of Cultural Narratives Shaping Emotional Responses to Infertility. IEEE, 2025.  
[ieeexplore.ieee.org/document/11167416](https://ieeexplore.ieee.org/document/11167416)
- Analysis of Deep Learning Algorithms on COVID-19 Radiography Database. IJAST.  
<https://github.com/amanjaiswal777/Analysis-of-Deep-Learning-algorithms-on-COVID-19-Radiography-Database>
- AUTHEER: A Voice-Based Speaker Authentication System. IJAST.  
<https://github.com/amanjaiswal777/AUTHEER-A-Voice-Based-Speaker-Authentication-Systems>