Muhammad Ahmer Saleem

AI Systems Architect · Staff / Principal AI Engineer
Agentic Systems · RAG & Retrieval · Multimodal AI · Computer Vision
ahmer@nixvix.ai · Lahore, Pakistan — Open to Global Remote · linkedin.com/in/ahmer-saleem

EXECUTIVE SUMMARY

AI Systems Architect building compound AI systems — agentic orchestration, distributed retrieval, and multimodal perception composed into enterprise platforms that hold up under real data, high concurrency, and strict cost curves. 11+ years and 80+ production deployments. Shipped the multi-agent LangGraph core running live on physical retail kiosks; engineered hybrid retrieval across a 3M+ cinematic asset library; and architected the computer vision engine behind a retail intelligence platform from founding to a P&G Partner Excellence Award across 200+ FMCG brands. Drawn to compliance-heavy, zero-error environments — including HIPAA clinical imaging and PHI de-identification — where an ungrounded output is a liability, not a rounding error.
Operating Principle: The model is rarely the bottleneck; compliance, latency, cost, and infrastructure are. I identify the single constraint that makes a system unshippable, design backward from it, and choose the model last.

CORE EXPERTISE

Agentic Systems & Orchestration: LangChain, LangGraph, CrewAI, State Machines (Supervisor/Routing Architecture), HITL Interrupts, Durable Memory Systems, Context Isolation.

Claude Platform & Tooling: Claude API (Prompt Caching), Claude Agent SDK, Model Context Protocol (MCP).

RAG & Enterprise Retrieval: ColPali (Patch-level MaxSim), ColBERT Late-Interaction, Hybrid Search (RRF), Contextual Chunking, RAGAS Triad Evaluation.

Multimodal AI & Vision: CLIP, SigLIP, DINOv2, LLaVA, Qwen-VL, Document AI (Layout-Aware Parsing, Coordinate Grounding), PyTorch, YOLO (v5-v12), SAM, ByteTrack.

Vector & Data Infrastructure: PGVector, Qdrant, Pinecone, Vespa, Weaviate, FAISS, HNSW/IVF Indexing, Multi-Tenant Namespace Isolation, High-Throughput Upsert Batching.

Infrastructure & MLOps: Modal, Replicate, AWS, GCP, Docker, Kubernetes, TensorRT, ONNX, Embedded Edge AI (NVIDIA Jetson, Coral TPU), Quantization (INT8/QAT).

Regulated AI Systems: HIPAA/GDPR Systems Design, PHI De-Identification Pipelines, Biomedical NER, DICOM Imaging, FHIR Standards.

Languages & Runtimes: Python, C/C++, CUDA, SQL, Bash.

PROFESSIONAL EXPERIENCE

Principal AI Architect & Founder · Nixense Vixion (AI Systems Consultancy)
Aug 2019 – Present

Global Enterprise Delivery · Expert-Vetted Top 1% Architect · Selected Talent, Upwork Executive Summit 2023

Featured long-term enterprise engagements:

Lead AI Engineer, Multimodal Search & Retrieval · Shotdeck
Mar 2025 – Present

CA, USA (Remote) · shotdeck.com · World's largest cinematic stills library — for directors, cinematographers, and advertising agencies

- **Built Shotdeck's end-to-end cinematic AI layer from a blank slate** — replacing rigid keyword lookup with a hybrid retrieval pipeline: vision-language embeddings, automated multi-label tagging, spatiotemporal video-clip extraction, and color search across a 3M+ image asset library.
- **Engineered an in-house multi-label classifier** that raised exact-match tagging accuracy from a 3% baseline to 71% (80–96% on individual fields across ~15 simultaneous tags/asset); proved empirically that a 20K similarity-curated training set beat a 100K random sample by 20+ percentage points.
- **Led a zero-downtime migration** from always-on GPU instances to serverless orchestration, cutting compute costs from ~$3.20/hr to ~$0.80/hr; brought embedding, color, and tagging models in-house within weeks during an incumbent vendor deprecation with zero user-facing disruption.

AI System Architect, Multi-Agent Orchestration · Biscuit AI
Apr 2024 – Mar 2025

CA, USA (Remote) · biscuitai.com · Funded retail AI platform running in production on physical kiosks — live customer interactions against real inventory and purchase decisions.

- **Shipped a production LangGraph state-machine core** — a supervisor node for dynamic intent recognition and routing across specialized sub-agents (inventory, text-to-SQL), with fallback control loops returning to the supervisor for human-in-the-loop (HITL) clarification on unresolvable ambiguity.

- **Designed a multi-stage evaluation framework** on RAG Triad metrics (groundedness, context relevance, answer relevance), wired into LangSmith tracing and production feedback loops — replacing subjective assessment with a rigorous regression gate.
- **Architected a dual-layer memory system** separating transient session state from a durable, cross-session preference layer distilled continuously from historical interactions; built the hub-and-spoke layer for structured extraction from unstructured retail data.
- **Recruited by the founder as primary architecture decision-maker** to lead 5 engineers through critical RAG gaps; isolated and fixed root-cause retrieval defects — rigid self-query metadata filters, numeric tool-parameter exceptions, and a misdiagnosed metadata-grounding bug — unblocking launch with immediate, client-confirmed gains.

Founding AI Architect, Retail Shelf Intelligence · Shelfr
Sep 2019 – Jan 2025

Dubai, UAE (Remote) · shelfr.ai · AI shelf intelligence and execution platform for global FMCG brands

- **Joined at founding as sole AI Architect to build a zero-to-production enterprise shelf-intelligence platform** scaling to 200+ global brands (P&G, Basamh, NFPC) across 5 years of continuous deployment; engineered end-to-end CV pipelines for SKU detection, planogram compliance, out-of-stock recognition, price tag extraction, and promotional displays, culminating in the **2025 P&G External Business Partner Excellence Award**.
- **Led the ML production lifecycle and a team of 4**; architected a per-client model-partitioning framework (treating multi-brand SKUs as a family of related problems) paired with geometric row-assignment layers to derive occupancy without auxiliary models, on auto-scaling GCP multi-GPU infrastructure.
- **Instituted automated, versioned mAP regression testing loops** combined with cross-SKU false-positive matrix analysis to catch misclassifications between near-identical product variants before they reached live production.
- **Engineered real-time out-of-stock detection and instant shelf-replenishment alerting loops**, shifting retail operations from delayed manual audit cycles to immediate, same-visit remediation; drove a documented +22% on-shelf availability and +25% shelf-compliance gain across global client deployments.

Generative AI Engineer · AI Humans Platform · Synthesys AI Studio
Jul 2021 – Jun 2023

London, UK (Remote) · synthesys.io · AI content platform generating video, voice, and image content at scale

- **Architected the foundational Generative AI content stack from the ground up** — production pipelines for photo-realistic talking avatars, lip-sync (Wav2Lip, DeepFaceLab), voice cloning (Tacotron2, Coqui, Tortoise), and text-to-image (Stable Diffusion); scaled the core fabric to 2M+ annual API calls.
- **Owned the voice training and deployment lifecycle end-to-end**: data-collection strategy, automated preprocessing, and custom evaluation loops turning raw audio into production-ready cloned voices at scale.
- **Led a migration to serverless GPU execution (Modal, Replicate) across a hybrid cloud (AWS, GCP, Paperspace)**; cut container images ~90% (68GB→7GB) and reduced annual compute spend ~80% from a $100K+ baseline while maintaining 99.9%+ uptime.
- **Engineered the robust REST API layer and third-party integration architecture** that established the platform's core infrastructure, successfully enabling seamless enterprise-scale content delivery and paving the way for downstream corporate partnerships, including integration into Canva.

ML Engineer · Clinical Echocardiogram AI Research · Edwards Lifesciences
Aug 2019 – Apr 2020

USA (Remote) · edwards.com · Research collaboration on clinical-grade automated echocardiogram analysis

- **Co-engineered a HIPAA/GDPR-compliant DICOM de-identification system** enabling large-scale clinical research: a multi-stage pipeline (Mask R-CNN FPN+ResNet-101 with temporal stabilization, EAST text detection) masking PHI outside the echo cone while preserving internal color-Doppler overlays.
- **Built an attention-based Multiple Instance Learning framework** for weakly labeled echo studies (variable bag sizes/frame counts), surfacing diagnostically relevant frames without frame-level labels — 91% test accuracy via a staged-difficulty curriculum and a ConvLSTM→VGG-19-with-attention-pooling redesign.
- **Reimplemented and validated the EchoNet-Dynamic pipeline end-to-end** across the Stanford and CAMUS datasets (~10,000 videos) for clinical heart-failure prediction; utilized DeepLabv3 frame-level LV segmentation and (2+1)D ResNet3D spatio-temporal regression to reproduce published benchmark performance (MSE 141.7, MAE 8.7).
- **Engineered a beat-by-beat cardiac function assessment engine** that derived automated End-Systole and End-Diastole frame identification through segmentation mask sequence analysis; diagnosed and remediated a critical class-imbalance tracking collapse that had been producing deceptive validation metrics, restoring reliable model evaluation for the clinical research team.

Additional Engagements:

RAG Engineer | Mizan (2025–2026) Engineered a solo, end-to-end citation-strict RAG pipeline for a premier law firm, parsing ~15M tokens (47K+ chunks) via a resumable, idempotent Docling and Qdrant ingestion engine with layout-aware table-splitting mitigations; deployed a strict Claude-powered verification layer enforcing `[source, heading path, page]` citations and multi-path isolated observability alongside an SSE-streaming stateless API.

Senior AI Engineer | Confidential HealthTech (2024 – 2025) Developed a HIPAA-compliant biomedical NLP pipeline achieving 85%+ PHI token-classification accuracy across unstructured clinical notes and structured lab tables. Designed a multimodal document retrieval layer using late-interaction architectures (ColPali/ColBERT-style) to preserve fine-grained token-level medical context over dense embeddings.

AI Engineer | Confidential Financial Institution (2024 – 2025) Architected a real-time ATM fraud-detection pipeline gating a fine-tuned TSN/ResNet-50 video action classifier behind lightweight person detection — so GPU cost tracks activity, not uptime — distributed across Celery/Redis workers and piloted against live camera feeds. Basis of a filed provisional patent for hybrid edge/cloud fraud detection.

CV Engineer | Fluffy Pet Technologies (2024 – 2025) Designed a production, queue-driven YOLO + ByteTrack service for collarless multi-dog event detection, with a custom tracklet-merging layer that fuses fragmented tracks of the same event and filters spurious short ones — resolving heavy occlusion and re-entry without collars or manual tagging.

CV Engineer | FOO Technologies (2022 – 2023) Built a modular identity-verification suite: facial liveness/PAD combining a VGG-19 spoof classifier with landmark-driven challenge-response (blink, wink, smile, head turn) at 90%+ accuracy, and LBP/SVM document tampering detection at 91%. Designed retention around derived depth masks rather than users' photos — minimizing biometric data held while still enabling model refinement.

CV / Edge AI Engineer | Canaryaware (2020 – 2021) Sole CV engineer on a retrofit forklift AI camera system detecting pedestrians and equipment proximity in real time — SSD/MobileNet quantized via QAT for Google Coral TPU, Intel Movidius, and Raspberry Pi, holding 30 FPS on-device across arbitrary camera placements under low light, occlusion, and motion blur. Detections feed the product's near-miss safety analytics.

Senior AI/ML Engineer · Xavor Corporation
Mar 2018 – Sep 2019

Lahore, Pakistan · Enterprise software company serving global technology clients

- Took enterprise-client ML/CV prototypes to production across multiple verticals; built automated data collection, filtration, and augmentation pipelines and shipped ML and IoT workloads with custom performance telemetry.

CV & Deep Learning Engineer · Intagleo Systems
Nov 2016 – Feb 2018

Lahore, Pakistan

- Replaced sensor-based detection systems with CNN-based visual detection in C++ and Python — automated shoe sizing at 0.2–0.3 inch tolerance, body segmentation for clothing-size estimation across 6 measurement entities at 80% accuracy.

Computer Vision Engineer · Techknox Systems
Nov 2015 – Nov 2016

Lahore, Pakistan

- Built multi-task deep learning models for retail visitor analytics — expression, gender, age, and clothing attribute recognition from live video.

EDUCATION

Bachelor's in Electrical & Electronic Engineering · University of Bradford, UK
2011 – 2015

Bradford, United Kingdom · Best Student Award · Best Final Year Project Award

CERTIFICATIONS

- Building with Claude API, MCP, Agent Skills & Subagents — Anthropic · Long-Term Agentic Memory in LangGraph · Multi-Agent Systems — CrewAI

References, portfolio links, and NDA-protected project details available upon request · ahmer@nixvix.ai