

Ali Ameen

AI Engineer | Production AI Systems | RAG & Multi-Agent Pipelines | Evaluation & Reliability

+92 331 414 2849 | aliameenrana@gmail.com | Lahore, Pakistan (open to remote / relocation)

[linkedin.com/in/aliameenrana](https://www.linkedin.com/in/aliameenrana) | github.com/aliameenrana | aliameen.com | aliameen.com/blogs

PROFESSIONAL SUMMARY

AI engineer building production Retrieval-Augmented Generation (RAG) systems: agentic pipelines, citation-grounded retrieval, and on-device machine learning (ML). Winthrop, which I built end to end, runs today at a private equity firm managing over \$2B, cutting investment memo turnaround from days of analyst work to 10 minutes by catching hallucinations before delivery. Also built two citation-grounded RAG platforms serving hundreds of thousands of documents at ~2.5s latency. Before AI, a decade shipping real-time systems under hard latency budgets on projects that raised \$15M+ and reached 200M+ viewers.

SKILLS

- **AI Engineering:** Agent orchestration and tool use, MCP (Model Context Protocol), LangGraph, multi-agent pipelines, RAG (Retrieval-Augmented Generation) and hybrid retrieval, forced tool calling, structured outputs, prompt engineering and caching, fine-tuning, embeddings, Anthropic API, AWS Bedrock, Groq, PyTorch, LangChain
- **Evaluation & Reliability:** Large Language Model (LLM)-as-judge, cross-family adversarial evaluation, hallucination detection, prompt-injection defense, groundedness and faithfulness scoring, span-level citation attribution, deterministic validators, error taxonomies, regression benchmarking
- **Backend & Infrastructure:** Python, FastAPI, asyncio, Postgres / pgvector (vector database), Redis, Docker, AWS (Bedrock, S3, EC2), Cloudflare Workers, Cloudflare D1, Nginx, Git
- **Real-Time & On-Device:** On-device machine learning (ML) inference, TFLite, 6DoF pose estimation, GPU/CPU profiling, Unity, Unreal, C#, C++

AI ENGINEERING EXPERIENCE

AI Engineer, Winthrop

Aug 2025 – Present

Production diligence system deployed at a private equity firm managing over \$2B. Turns 40 to 175 page deal documents into structured investment memos where every claim links back to its source, cutting turnaround from days of analyst work to 10 minutes and saving hundreds of thousands a year in analyst fees, with per-document cost held down by prompt caching over shared document context.

- Designed the orchestrator: each stage runs as a tool-scoped agent, validated against a JSON Schema contract before the next stage dispatches. State checkpoints after each stage, so failed runs resume without reprocessing the document.
- Wrote the ingestion layer: four PDF libraries run in parallel — pdfplumber for layout-aware text and tables, PyMuPDF for rasterization, pypdf and pymupdf4llm as fallbacks — plus python-docx and python-pptx for DOCX and PPTX.
- Engineered the claim-verification layer: every claim is page- and region-linked to its source, normalized, and matched for a fabricated / wrong-page / out-of-tolerance verdict. Four more checks — vision review of charts and scans, sandboxed arithmetic re-derivation, a source-coverage sweep, and an HTML structural pass — catch hallucinations the text match misses. Together: 87% of unsupported claims caught before delivery, across 150+ documents.
- Ran an adversarial audit loop: Claude Opus 4.8 and GPT-5 cross-review each memo against its source for numerical, analytical, and logical errors, grouped into error classes. No prompt or schema change ships until re-tested against a fixed document set. Cut the most common error classes by up to 92%.
- Deployed on the client's own Ubuntu servers behind a provider abstraction: AWS Bedrock (via boto3, with Guardrails) primary, direct Anthropic SDK as fallback when Bedrock throughput limits bind. FastAPI/Uvicorn, SQLAlchemy 2.0, Alembic, Redis-backed rate limiting and spend caps, Nginx, systemd.

AI Engineer, [Infin8](#) (UGC Marketplace)

May 2026 – Aug 2026

Creator-brand marketplace where brands run campaigns and select creators, and creators submit content against campaign briefs. Contributed the onboarding, campaign workflow, chat, and content moderation systems.

- Modeled onboarding and campaign creation as LangGraph state machines rather than free-form agents, since both are bounded, step-based flows: each step persists through tool calls, with users free to jump back and edit any prior step without restarting.

- Extended the chat layer site-wide so any UI action — campaign creation, submissions, analytics, payments — is also available conversationally through tool use.
- Routed content moderation by cost tier rather than one model for everything: cheap, fast models screen plain text, a stronger tier handles PDF and document parsing, and Kimi handles image and video, each checked against campaign brand guidelines before publishing.

AI Engineer, Claude-e-Azam & Verbatim (aliameen.com/verbatim)

May 2026 – Aug 2026

Two production citation-grounded RAG (Retrieval-Augmented Generation) platforms. Claude-e-Azam answers as Muhammad Ali Jinnah using only his documented words; Verbatim extends it to Gandhi, Einstein, and Darwin, each citing every substantive claim. Hundreds of thousands of documents, ~2.5s latency, on Cloudflare Workers with Hyperdrive, behind a provider abstraction across Anthropic, Groq, and Gemini.

- Fused hybrid dense and keyword retrieval via Reciprocal Rank Fusion over pgvector and Postgres full-text search, with an LLM query-classification stage routing biographical, modern-topic, and follow-up questions down distinct retrieval paths.
- Isolated data per persona: separate Postgres tables per figure with halfvec embeddings (~50% storage reduction) and ivfflat/exact-scan ANN indexing; multilingual support (Roman Urdu/Hindi) via pre-retrieval query translation; streaming-safe per-claim citation resolution.

AI Engineer, [ATS Resume Review](#) (MCP Server)

Sep 2026 – Present

- Built an MCP (Model Context Protocol) server so AI agents can run resume/ATS reviews: parses a resume into a deterministic ATS score plus Groq feedback. Since an LLM reads untrusted resumes, output is JSON-schema validated as a guardrail against prompt injection, with D1-backed rate limiting.

Technical Lead, Real-Time Edge Pose Tracking ([Top Spot Games](#))

Mar 2023 – Jul 2025

Real-time 6DoF pose-tracking ML pipeline running fully on-device on low-end Android hardware, with no depth sensor.

- Directed end-to-end development (dataset creation, annotation, model training, on-device inference) for real-time bat and body tracking from the front camera, optimized to run within the compute and battery budget of low-end Android phones.
- Led a cross-functional team of ML and Unity engineers, owning the split between what ran in the model and what ran in the game runtime, and the annotation and evaluation process that kept tracking accuracy stable as the training set grew.
- Extended the model into a modular tracking toolset for Wii Sports-style gameplay across cricket, baseball, tennis, and table tennis; drew interest from ICC, ECB, MLB, and Wimbledon and contributed to a \$5M seed raise.

AI Engineer, [AINT](#) (Personal Project)

Mar 2026 – May 2026

- Fine-tuned GPT-2 (small) on a synthetic corpus for a consistent comedic voice, paired with instruction-conditioned generation so it produces new on-topic lines at inference rather than replaying training data. Shipped to Google Play; live corpus of 5,800+ generated sentences across 117 topics.

GAME DEVELOPMENT & SYSTEMS BACKGROUND

A decade shipping performance-critical software under hard latency budgets before moving to AI.

- **[Centurion VR](#), Senior Developer / Technical Lead (Nov 2018 – Feb 2023):** Built the high-fidelity VR cricket simulation broadcast live by Star Sports India to 200M+ viewers. Sub-frame bat-ball collision physics, strict VR frame-rate guarantees, and a real-time AWS backend (DocumentDB, S3, Node.js) for live delivery data. Won the contract over Accenture on technical merit and helped secure \$10M investment.
- **[OZI Technology Group](#) & [Sablo Studio](#), Unity Developer (Oct 2017 – Oct 2018):** Led two developers and two designers to ship a Homescapes-style renovation game feature-complete in six months at OZI, Pakistan's largest game group, cutting development timelines by ~20% across its studios. At Sablo, a Google-partner studio with millions of downloads, shipped a farm simulation and a multiplayer paintball FPS with real-time networking on low-end hardware.
- **BugDev Studios (now [Conovo Technologies](#)), Unity Developer (Aug 2016 – Sep 2017):** Developed multiplayer VR gameplay for Curse of the Sunken Temple using LEAP Motion for hand tracking, holding a stable 90 FPS for user comfort.

EDUCATION

B.S. Electrical Engineering, LUMS — full merit scholarship (PKR 8M)

2010 – 2014