

Ali Ameen

AI Engineer | Data & ML Engineering | RAG & Multi-Agent Pipelines | Evaluation & Reliability

+92 331 4142849 | aliameenrana@gmail.com | Lahore, Pakistan (Open to remote or relocation)

linkedin.com/in/aliameenrana | github.com/aliameenrana | aliameen.com | aliameen.com/blogs

PROFESSIONAL SUMMARY

AI engineer building production multi-agent and RAG systems. At Winthrop, built the AI platform a \$2B+ private equity firm uses for deal analysis, cutting memo cost from \$10,000+ to about \$5 and saving \$2.5M+. Before that, 8 years in real-time systems and ML, including predictive models used on Star Sports' live IPL coverage.

SKILLS

- **AI Engineering:** Multi-agent systems, LangGraph, LangChain, MCP, RAG, hybrid search, embeddings, fine-tuning, Anthropic API, AWS Bedrock (Knowledge Bases, Guardrails), PyTorch
- **Evaluation & Security:** LLM-as-judge, hallucination detection, guardrails, prompt-injection defense, citation grounding
- **Data & Backend:** Python, FastAPI, Postgres/pgvector, Redis, Kafka, Snowflake, AWS (S3, EC2), Azure, Docker, Cloudflare
- **ML & Real-Time:** Predictive modeling, pandas, scikit-learn, Jupyter, SageMaker, TFLite, pose estimation, C++, C#

AI ENGINEERING EXPERIENCE

AI Engineer, Winthrop

Jul 2024 – Present

- Built a multi-agent system turning 40–175 page deal documents into source-linked memos in 10 minutes instead of days.
- Cut the cost of one memo from \$10,000+ in analyst hours to about \$5, saving the firm \$2.5M+ across 270+ documents.
- Built a citation-grounded analyst chatbot and agents that pull exact historical insights from company financial reports.
- Caught 87% of unsupported claims before delivery; cross-model audits cut the most common errors by up to 92%.
- Runs on the client's servers with Bedrock Guardrails and prompt-injection defense; aligned with SOC 2 and ISO 27001.

Client: [Infin8](#) (UGC marketplace)

May 2026 – Aug 2026

- Built onboarding and campaign creation as LangGraph agents, letting users edit any earlier step without starting over.
- Made every site action (campaigns, submissions, analytics, payments) available through chat using tool calls.
- Set up tiered moderation: cheap models for text, stronger ones for documents, a vision model for images and video.

PERSONAL & OPEN-SOURCE PROJECTS

- [Claude-e-Azam & Verbatim](#): Cited RAG for Jinnah, Gandhi, Einstein, and Darwin; 100,000s of documents, 2.5s avg latency.
- [ATS Resume Review \(MCP Server\)](#): Lets AI agents score resumes, with schema-checked output to block prompt injection.
- [AIN'T](#): Fine-tuned GPT-2 for a consistent comedic voice; live on Google Play with 5,800+ generated lines across 117 topics.

SYSTEMS BACKGROUND

Technical Lead, [Top Spot Games](#)

Mar 2023 – Jun 2024

- Led ML and Unity engineers building real-time 6DoF bat and body tracking on low-end Android phones, no depth sensor.
- Owned the full ML pipeline: synthetic dataset creation, annotation, model training, and on-device inference.
- Drew interest from ICC, ECB, MLB, and Wimbledon, and helped raise a \$5M seed round.

Tech Lead / Senior Developer, [Centurion VR](#)

Nov 2018 – Feb 2023

- Built predictive ML models with feature engineering on bowler and team analytics each IPL season to decode game plans.
- Ran live broadcast analytics for 200M+ viewers; the segment drew 1.5x the TRP of other analysis segments.
- Designed the data pipeline: live ball-by-ball feeds via Kafka into S3, DocumentDB, and Snowflake, with data back to 2006.
- Led the real-time cricket simulation for Star Sports India; beat Accenture for the contract and helped raise \$10M.
- [OZI Technology Group & Sablo Studio](#), Unity Developer (Oct 2017 – Oct 2018): Led a team of four to ship in six months, cutting timelines ~20%; built real-time multiplayer.
- [BugDev Studios \(now Conovo Technologies\)](#), Unity Developer (Aug 2016 – Sep 2017): Built multiplayer VR with hand tracking at 90 FPS.

EDUCATION

B.S. Electrical Engineering (Computer Science & Electronics), LUMS, full merit scholarship (PKR 8M)

2010 – 2014