

AYAAN KHAN

9343282801 | ayaan.khan2812@gmail.com | linkedin.com/in/ayaankhan2812 | github.com/ayaankhan28 | [portfolio](#)

EDUCATION

Birla Institute of Technology and Science, Pilani

B.S. in Computer Science (Hons)

Pilani, India

July 2023 – August 2027

ACHIEVEMENTS

- **1st place**, iQOO Hackathon 2026- built on device 3d physics simulation Engine to convert study material in to interactive content
- **2nd place** (solo) in the **Emerging Tech Hackathon** by Metadome.ai build local on device Image Search Engine.
- Press coverage across [Times of India](#), [Deccan Herald](#), [Business Standard](#), [EdexLive](#), and [Silicon India](#) for projects in CV, GenAI, and on-device ML.
- Contributed **5+ merged PRs** to **mcp_ctl**, an open-source MCP server scaffolding CLI by Runnable — improved flag parsing and developer UX.

EXPERIENCE

Software Engineering Intern

Omium.ai

April 2026 – May 2026

Remote — San Francisco

- Worked across a **16-service polyglot backend** (Python/FastAPI, Go/gRPC) on AWS EKS — processed **341 agent runs and 12,455 billing events** within 24 hours of launch.
- Built a **3-tier checkpoint system** (Redis → PostgreSQL → S3) with **<5ms write latency**, circuit breaking on gRPC outage, and async LZ4-compressed S3 offloading.
- Designed a **self-healing failure pipeline** using MAST failure classification, ELO-ranked multi-LLM fix generation, and automated GitHub PR creation — reducing manual intervention for agent crashes.

Forward Deployed Engineer

AiStudio.ae

June 2025 – Sep 2025

Remote — Dubai

- Partnered with enterprise customers and founders to translate ambiguous operational workflows into production AI systems, rapidly shipping features through continuous user feedback.
- Owned deployment of a **multi-agent enterprise assistant** integrating CRM, Salesforce, Excel, document ingestion, **Knowledge Graph RAG**, and vector search into customer workflows.
- Engineered agent infrastructure with **workflow orchestration**, **modular tool routing**, **async execution**, **persistent state**, and evaluation-driven retrieval across enterprise data sources.

Software Engineering Intern

Scaler.com

June 2024 – Feb 2025

Bangalore, India

- Built an **AI support chatbot** on FastAPI with RAG-based retrieval — cut GPT API calls by **25%** and reduced human ticket volume by **47%**.
- Designed a **multi-model orchestration layer** with centralized tracing via Langfuse, coordinating 3+ LLMs across support and interview workflows.
- Implemented **WebSocket-based live interview sessions** with proctoring and speech-to-speech interaction across 45–60 min sessions.
- Built async notification pipelines using **Celery + Redis**, dispatching WhatsApp and email events at scale.

PROJECTS

ShellBreaker.com — Interactive Linux Learning Platform | Bubblewrap, tmux, cgroups v2, FastAPI, Redis

Oct 2025 – Present

- **Problem:** No platform existed for practicing real Linux commands in isolated environments without risking the host system.
- Built a **multi-tenant namespace-isolated sandbox** using Bubblewrap (**<8 MB/session**), tmux, and cgroups v2 with CPU/memory/PID isolation and **<2s** automatic termination on limit breach.
- Cold-start under **10s; 300+ syscalls blocked** via seccomp; automated GC reduced orphaned sessions by **95%**; PostgreSQL audit logging captures **99%** of commands.
- **Supports 70+ concurrent sandboxes** on a 4 GB VM; **1,400+ users** have executed **5,700+ Linux commands** across **1,000+ practice sessions** (avg. **10 min/session**)

IRIS Studio — Multi-Agent Execution Platform [Live Demo] | Python, FastAPI, LangGraph, Redis, PostgreSQL, MCP

May 2026 – Present

- Built a **LangGraph-based agent execution engine** with supervisor routing, conditional branching, feedback loops, human approval checkpoints, and persistent state across workflow runs.
- Engineered a **custom LLM observability SDK** with ContextVar decorator-based span instrumentation and a non-blocking **Redis Streams ingest pipeline** (consumer group + DLQ) mirrors OpenTelemetry trace hierarchy with per-provider cost estimation and **PII** redaction.
- Integrated GitHub, Slack, Gmail, Notion, and Telegram; agents execute multi-step workflows end-to-end with real-time tracing through a FastAPI analytics backend.
- Published the **iris-ai-cli** AI CLI on PyPI with **6.6k+ downloads**.

PhotoSage — On-Device Semantic Image Search [App Store] | Swift, Python, CLIP, FAISS, USearch

Dec 2024 – Mar 2025

- **Problem:** Searching through thousands of photos on mobile requires manual scrolling with no semantic understanding.
- Built a **fully on-device** search engine using CLIP embeddings and FAISS/USearch — all inference runs locally with no data leaving the device.
- Matched or exceeded Google Photos on **70% of internal benchmark queries; 200+ downloads** on the App Store.

SKILLS

Languages: Python, Go

Frameworks: FastAPI, LangGraph, Celery

Databases: PostgreSQL, Redis, MongoDB

Protocols: gRPC, WebSockets, SSE, MCP

Architecture: Microservices, Multi-Tenant Systems, Agent Orchestration

Observability: OpenTelemetry, Langfuse, Redis Streams, Structured Logging

Cloud/DevOps: AWS (EC2, S3, EKS), Docker, Kubernetes, Kafka, CI/CD

ML/AI: CLIP, FAISS, RAG, Knowledge Graphs, Vector Search