

Haris Jalal

APPLIED AI & AGENT PLATFORM ENGINEER

San Francisco, CA | harisjalal502@gmail.com | +1 (267) 601-6614 | linkedin.com/in/haris-jalal | github.com/harisjalal502

Applied AI and agent-platform engineer with 4+ years building governed production agent systems, enterprise RAG, and full-stack AI infrastructure at Oracle. Turns brittle prototypes into reliable, observable platforms spanning orchestration, retrieval, evaluation, safety, and deployment, supporting 60+ workflows, 1.2M documents, 8.6M indexed chunks, and 50K+ monthly conversations.

EXPERIENCE

Oracle Corporation | Software Engineer II (AI Agent Platform)

Feb 2025 - Present | San Francisco, CA

- Built ArtiFoundry, a governed MCP/A2A platform with a schema-driven catalog and visual workflow editor that cut new-assistant setup from weeks to under 10 minutes.
- Automated CI/CD, security and policy gates, Kubernetes rollout, and Agent Gateway registration, cutting production release time from three days to 45 minutes.
- Designed a LangGraph runtime for 20+ production assistants and 60+ workflows; dynamically compiled typed graphs and published workflow-state events via Redis Streams across 30 Kubernetes pods, enabling hot updates without application redeployment.
- Implemented two-stage RAG over 1.2M enterprise documents and 8.6M indexed chunks with Oracle AI Vector Search and Cohere Rerank, improving retrieval quality by 12 percentage points and citation accuracy from 82% to 93%.
- Benchmarked Matryoshka embeddings from 1,536 to 512 dimensions, reducing vector-index size 63% and p95 retrieval latency from 410 ms to 285 ms with less than a one-point retrieval-quality tradeoff; added versioned synchronization and source-level retrieval traces.
- Hardened inference with Presidio/spaCy PII redaction, schema-validated retries, guardrails, and health-aware provider failover, reducing user-facing LLM failures 41% and maintaining 99.95% monthly availability through six upstream-provider incidents.
- Mentored four engineers through structured onboarding, architecture walkthroughs, design reviews, and paired production launches; all four independently shipped production changes within 60 days, cutting median ramp time from eight to four weeks.

Oracle Corporation | Software Engineer (Conversational AI Platform)

Jul 2022 - Jan 2025 | San Francisco, CA

- Built and iterated the conversational AI foundation serving 50K+ monthly conversations across 10 Oracle business units, increasing containment (sessions without live-agent escalation) from 49% to approximately 65% through retrieval, content, and workflow improvements.
- Developed a Preact/TypeScript SDK adopted by 18 internal applications, enabling assistants through one JavaScript include with streaming citations, OAuth, multilingual and voice input, and WebSocket live-agent handoff.
- Built a Java 21/Helidon ingestion control plane synchronizing 2.8M objects across 14 enterprise sources, reducing stale or failed synchronizations 68% and median diagnosis time from 90 to 18 minutes via per-object ETags and explicit state transitions.

Drexel University | ML Research Engineer

Apr 2021 - Sep 2021 | Philadelphia, PA

- Built scikit-learn and Prophet forecasting models that reduced sales-forecasting error below 5% across seven product lines, outperforming manual planning baselines.

Oracle Corporation | Software / QA Engineer

Apr 2020 - Sep 2020 | Philadelphia, PA

- Shipped four customer-facing React features and cut Electron startup time 14% through performance profiling and optimization.

Merck | Business Analyst

Apr 2019 - Sep 2019 | Philadelphia, PA

- Managed 500+ requirements for an enterprise Veeva CRM rollout and automated Salesforce/VBA reporting for cross-functional business teams.

SELECTED PROJECTS

AutoArena | Autonomous Adversarial Research System | Antler Hackathon

- Built a Python adversarial-evaluation system with evidence-gated experiments, deterministic JSONL replay, immutable leaderboard receipts, full trace export, and lineage-aware rollback; increased strategy win rate from 44% to 69% across 112 held-out evaluations on Modal.

Better Call Ron | Real-Time Voice Agent with Guardrails | Y Combinator Hackathon

- Built a guardrailed FastAPI voice agent for caller triage and live attorney transfer, achieving a 74-second median completion time and 94% success across 50 end-to-end trials; implemented signed webhooks, policy-bounded tool calls, attorney matching, and structured SMS/email briefs.

Tarry | Embodied Multimodal Memory Agent | Y Combinator Hackathon

- Built a Reachy Mini multimodal memory agent with OpenAI Realtime and GBrain that converted live audio and whiteboard images into meeting notes and persistent office memory, achieving 90% top-1 retrieval accuracy across 120 follow-up memory queries.

PROFESSIONAL SERVICE

Peer Reviewer: NeurIPS 2026 and COLM 2026

TECHNICAL SKILLS

Languages: Python, Java, TypeScript, SQL

AI & Agents: LangGraph, retrieval-augmented generation (RAG), LLM evaluation, Model Context Protocol (MCP), Agent2Agent (A2A), tool calling, OpenAI Realtime, Oracle AI Vector Search, Cohere Rerank, Matryoshka embeddings

Backend & Infrastructure: FastAPI, Helidon, Redis Streams, REST APIs, WebSockets, OAuth, Kubernetes, Docker, OCI, CI/CD, Modal

Safety, Observability & Frontend: Presidio, spaCy, LLM guardrails, observability, React, Preact

EDUCATION

Drexel University | Bachelor of Science in Computer Science

Sep 2017 - Jun 2022 | Philadelphia, PA

Minors: Economics and Mathematics | Honors: Dean's Scholarship