Iljard Xhani

AI Application Engineer (LLM + RAG + Reliability)

xhani.iljard@gmail.com

SUMMARY

AI Application Engineer focused on shipping production-ready LLM features with measurable quality gates. Builds end-to-end systems across retrieval, orchestration, evaluation, and operational reliability. Portfolio includes six detailed case studies across RAG, multi-tenant SaaS controls, and workflow automation.

CORE SKILLS

- LLM app engineering: prompt contracts, tool/function calling, deterministic APIs
- RAG systems: ingestion, chunking, hybrid retrieval, reranking, citation grounding
- Evaluation and release safety: faithfulness checks, regression gates, CI evals
- Reliability: observability, rate limits, fallback paths, audit logging
- Backend: FastAPI, Python, Pydantic, Postgres, pgvector, Docker, n8n
- Architecture: multi-tenant auth, policy enforcement, runtime guardrails

SELECTED PROJECTS

RAG Evaluation Lab (Release Gate System)

- Built policy-driven evaluation orchestration to block low-quality releases before deployment.
- Implemented thresholds: max quality regression -0.02, min pass rate 0.90, min faithfulness 0.88, max p95 latency 1900ms.
- Added PR smoke checks plus nightly benchmark runs for faster feedback and long-horizon quality tracking.

DocChat RAG (Production)

- Built PDF/docs ingestion-to-answer pipeline with hybrid retrieval and citation-constrained responses.
- Added runtime guardrails, fallback lanes, and retrieval confidence checks to reduce unsupported answers.
- Instrumented faithfulness, fallback-rate, and p95 latency signals for production monitoring.

Mini SaaS: RAG for Teams

- Built multi-tenant AI platform layer: auth context, org isolation, plan entitlements, quota and rate controls.
- Implemented usage metering, billing flow hooks, and structured audit events for governance.
- Validated tenant isolation and burst-traffic rate-limit behavior under plan policies.

Customer Support AI Triage (n8n + FastAPI)

- Designed event-driven triage workflow for email/DM classification, draft generation, routing, and ticket creation.
- Enforced human-in-the-loop approvals and event-level auditability for outbound responses.
- Added idempotent retries, dead-letter handling, and low-confidence fallback templates.

PROFESSIONAL FOCUS

Open to AI Application Engineer roles building stable, measurable, and production-oriented AI features.