MUDIT GARG

Backend Engineer

Delhi, India | +91 7042466813 | muditgarg.tech@gmail.com | muditgarg.xyz

linkedin.com/in/muditgargdev | github.com/MuditGarg007 | leetcode.com/u/muditcodez

SUMMARY

Backend engineer who designs and ships production-grade, scalable systems end to end. Pairs strong backend fundamentals in high-performance APIs, distributed data and reliability at scale with hands-on applied AI across RAG, multi-agent systems and LLM evaluation, delivered on cloud infrastructure with modern DevOps practices.

EXPERIENCE

OffplanWise (Anav Consulting) · Full Stack Developer Intern · *Remote, Dubai, UAE* **Apr 2026 – Jul 2026**

- Owned the FastAPI + PostgreSQL backend (~36K LOC Python, 81 endpoints, 98 Alembic migrations, 1,100+ tests) for a production RAG chatbot at offplanwise.com serving **10,000+ monthly queries across 20+ countries** over 20,000+ documents.
- Designed dual-KB hybrid retrieval: 4 parallel legs (pgvector HNSW + Postgres FTS over curated and scraped corpora) fused via RRF, one Cohere rerank and a 70/30 curated-first context quota; lifted recall from **62% to 86%** on a 131-question eval set.
- Built a deterministic turn orchestrator (intent classifier, response-shape routing, Redis memory, SSE streaming) plus SQL lanes for market stats and DLD transaction trends; **~84% faithfulness, ~90% citation accuracy** via LLM-as-judge + Langfuse.
- Shipped 3 fault-tolerant ingestion pipelines (news scraper over 9,400+ URLs, YouTube transcripts, 12-stage listings ETL over 4 portals) with resumable checkpoints and retry budgets; **500+ docs/week**, staleness cut from 7 days to <24 h.
- Cut the listings clean stage from **78 min to 40 s (117x)** by batching ~50K per-row DB round-trips; cut retrieval latency **45%** and LLM spend **~35%** via Redis caching, pooled parallel queries, prompt caching (77% hit rate) and context truncation.
- Secured the API with Google OAuth, short-lived JWTs + rotating refresh tokens and 4-layer guardrails (input validation, Lua-scripted Redis rate limiting, streaming moderation, output checks); lead-capture routing lifted lead conversion **30%**.

PROJECTS

AutoSploit AI · *LangGraph, NestJS, Kubernetes, gVisor, Redpanda, Vault* **Live Demo · Code**

- Built a platform that deploys a user's GitHub repo into a sandbox and points an autonomous LangGraph attack agent at it, streaming findings and cost over SSE; landed **real exploits on OWASP Juice Shop** (SQLi, admin JWT login, order injection).
- Architected a 5-service Bun + Turborepo monorepo (Next.js client, NestJS control plane, 3 Python services) on PostgreSQL, BullMQ/Redis, Redpanda (Kafka), Vault Transit and S3, with bounded-context data ownership.
- Enforced multi-tenant isolation with per-engagement K8s namespaces on gVisor, default-deny egress (**exactly 2 allowed edges**) and a fail-closed interceptor checking every tool call against a scope allowlist and spend cap.

Prism · HTTP Reverse Proxy · *Go, net/http, Prometheus, Docker* **Code**

- Built an nginx-style reverse proxy in Go (~25K LOC) with no httputil.ReverseProxy: hand-written forwarding, streaming, WebSockets and safe retries, 4 load balancers (P2C least-conn, consistent hash), GCRA rate limits and lock-free hot reload.
- Added per-backend circuit breakers, outlier ejection and a retry budget that held load amplification to **1.12x (vs 3.0x)** in a full outage, plus an RFC 9111 cache whose request coalescing collapses **1,000 concurrent requests into 1 backend fetch**.
- Benchmarked against nginx 1.28: **68.8K req/s** proxied (~70% of nginx), **99K req/s** on cache hits (level); cut hot-path allocations up to 31% via profiling; 361 tests + 9 fuzz targets, all under -race.

TECHNICAL SKILLS

Languages: Python, Go, TypeScript, JavaScript, Java, SQL
Backend & APIs: FastAPI, NestJS, Node.js, Express, REST APIs, Server-Sent Events (SSE), event-driven microservices
Databases & Messaging: PostgreSQL (pgvector), Redis, MongoDB, Kafka / Redpanda, BullMQ, Drizzle ORM
AI/LLM: RAG, LangGraph, LangChain, MCP, multi-agent orchestration, LLM evaluation (Langfuse, LLM-as-judge), PyTorch
DevOps: Docker, Kubernetes, Helm, gVisor, Terraform, GitHub Actions, AWS, OpenTelemetry, Prometheus, Grafana, Vault

ACHIEVEMENTS & CERTIFICATIONS

- **LeetCode:** 330+ problems solved; Knight badge, contest rating 1,955 (top 3.37%).
- **Hackathons:** 4th / 1,200+ teams (Mind Installers, May 2025); Top 25 / 2,500+ (HackXtreme, Apr 2026); Top 15 / 800+ (Hackstasy, Apr 2025); Top 15 / 850+ (HackFormers, Jan 2025).
- **HackerRank:** 5★ Gold (Python, Java).
- **Certifications:** NVIDIA Fundamentals of Deep Learning · Google Cybersecurity Professional · Anthropic Claude Code 101
- **Leadership:** Core Team, Microsoft Tech Community (Amity): mentored 100+ juniors; ran contests for 300+ participants.