Nikhil Yarram

yarramnikhil@gmail.com | (206) 465-1271 | Minneapolis, MN (Relocating to San Francisco, CA) | LinkedIn | Portfolio

SKILLS

Languages: Java, Python, Typescript, SQL

Backend & Architecture: Spring Boot, Kotlin, REST API, GraphQL, Microservices, Gatling, JUnit/Mockito, JPA Hibernate

Data & Messaging: PostgreSQL, NoSQL (MongoDB), Flyway, Redis, Kafka, OpenSearch/Elasticsearch

Cloud & Infrastructure: AWS (SNS/SQS, EC2, S3), Docker, Kubernetes, EKS, Terraform, CI/CD (GitHub, Jenkins)

AI & Tooling: RAG pipelines, LLM APIs, prompt/context engineering, agent frameworks, MCP (Model Context Protocol), LangChain, LLM-as-judge evals, AI-assisted development, Claude Code, Codex, Cursor

EXPERIENCE

Chewy **Minneapolis, MN**
Software Engineer *April 2025 – Present*

- Built an allocation-level fulfillment engine sourcing single SKUs across multiple FCs (partial shipments, floor denials); scaled to 100% of production, unlocking $2M in savings.
- Architected a Spring Boot EKS module bridging SNS to Kafka for real-time order notifications, implementing at-least-once delivery, order-keyed partitioning, and DLQ redrive; targets ~$2.2M in annual cost deflection.
- Engineered a zero-downtime V1-to-V2 order-event migration across 50 SNS subscriptions, designing a transactional outbox with stable order-ID bucketing to guarantee lifecycle schema consistency without coordinated cutovers.
- Developed an async payment-resiliency state machine via FIFO SQS/SNS and DLQs; decoupled Checkout from 3 synchronous gateways during 5xx outages to preserve order intake with a 3-day recovery SLA.
- Spearheaded remediation of a distributed tax defect, using Snowflake lineage analysis to reduce a 1.25M-line blast radius to 1,009 lines; architected an FC-keyed consistency path across 3 core services to resolve a $391K historical variance.

Verizon (via Indigo Systems) **Chicago, IL**
Software Engineer *September 2023 – April 2025*

- Built a distributed log-correlation engine (Spring Boot) over OpenSearch streams processing 100K+ daily audit events; iterative connected-components to stitch fragmented events into IAM journeys; cut debugging time 50%.
- Architected an LDAP identity-resolution module featuring a custom UnboundID connection pool with leak tracking and auto-rebinds; deployed via Docker/Kubernetes to support millions of monthly authentication operations.

Oshkosh Corporation **Oshkosh, WI**
Software Engineer Co-op *June 2022 – May 2023*

- Architected GraphQL API endpoints with FastAPI for a fleet management platform, optimizing real-time analytics by 20%.

PROJECTS

Litmus

- Built an open-source Python eval/observability harness for LLM pipelines: async eval runner, golden-dataset regression evals, calibrated LLM-as-judge scoring, OpenTelemetry tracing to Jaeger
- Added CI quality gates via GitHub Actions - hermetic keyless evals on every PR; litmus diff fails the build on regressions vs. a committed baseline
- Dogfooded on Astrodigest prod pipeline; caught bugs - OTLP exporter 404, MCP tool-calling SDK inheritance failure

AstroDigest - AI-curated weekly space & astronomy news digest

- Designed a provider-agnostic model-routing layer mapping relevance-score tiers to swappable LLM backends (Claude/Groq), decoupling core pipeline logic from single-provider lock-in.
- Built a Cloudflare Workers ingestion layer with idempotent upserts to ensure fault-tolerant daily runs.

EDUCATION

The University of Wisconsin - Madison | Bachelor of Science, *Computer Science* | May 2023