APURVA MODI

Seattle, WA | modiapurva03@gmail.com | LinkedIn | GitHub | +1-757-670-5100

TECHNICAL SKILLS

- **Languages:** Rust, Python, Java, JavaScript, TypeScript, HTML/CSS, PHP, SQL
- **Frameworks & Libraries:** React, NextJS, Node.js, Django, FastAPI, Spring Boot, Express, Keras, TensorFlow, Tailwind
- **Cloud & AWS Services:** ECS, Lambda, S3, SQS, DynamoDB, CloudFormation, CDK, IAM, KMS, Kinesis, AppConfig, ALB, CloudWatch, Bedrock, Q Business, AppFabric
- **Tools & Concepts:** Git, Docker, REST/SOAP APIs, Microservices, OAuth, CI/CD, A/B Testing, Distributed Systems, Leader Election, Tokio Async Runtime, MCP Servers, AI Tooling, Ollama, Vercel

PROFESSIONAL EXPERIENCE

Software Development Engineer II, Amazon – Bedrock Infrastructure *Mar 2025 – Present*

- Inference Scheduling Service (Rudi) primarily for Anthropic and Amazon Nova models:
  - Co-architected and developed Rudi, a Rust-based Bedrock inference scheduling service handling **1000+ TPS**; governs timing and ordering of inference requests from FES to model hosting layer — implementing concurrency-per-variant enforcement, priority-based queuing with weighted round-robin scheduling, request throttling/load shedding, and enabling Provisioned Throughput V2 without hard-allocating capacity
  - Built on single-threaded Tokio scheduling engine with MPSC/oneshot channel patterns for **microsecond-level** scheduling decisions; deployed across **all Bedrock regions**
- Priority Queue Migration:
  - Implemented two-phase **zero-downtime** migration from legacy 4-value (0,1,2,3) to standardized 7-tier priority system (P5–P100), with bidirectional session token normalization handling values 0–150 for rollback compatibility
  - Built configuration-driven priority weights, AppConfig/Rust fallback configs, and capacity allocation constraints; deployed across Beta, PreProd, and Prod with **zero incidents** over 6-week rollout enabling **2 new revenue tiers** (Premium OD, Best Effort OD)
- Quality of Service (QoS):
  - Implemented QoS tier mapping (RudiPriority → QoS 0/1/10) in schedule responses and forwarding logic for Anthropic models (Claude Sonnet, Haiku, Opus); added preemptable flag in capacity constraints and AppConfig-based model QoS detection
  - Built unified priority assignment avoiding separate QoS vs. non-QoS code paths; updated capacity allocator to deduct preemptable consumption only for QoS-supported models
- RAMA Classification Consolidation:
  - Led **4-team** consolidation expanding from 2-value to **10-classification** system (PTv2, Priority Access, OnDemand, Flex, Batch); implemented fail-closed exception throwing in Rudi and FES for unknown classifications with SEV2 alarms
  - Built thick client with AppConfig-based cached overrides (**50+** account-level imports), shadow mode validation framework, and session token migration including RAMA classification for request lifecycle consistency
- Wiring ServiceTier API for OpenAI gpt model lauch on Bedrock:
  - Implemented ServiceTier-to-RudiPriority mapping with configuration-driven rules engine, **15+ rules**) supporting FLEX/PRIORITY/DEFAULT tiers and RAMA BEST_EFFORT integration; built dual-operation mode with RequestType fallback and feature flags for gradual rollout
- Leader-Follower & Game Day:
  - Implemented leader-follower architecture with DynamoDB coordination, request forwarding, and broadcast mechanisms; automatic failover recovery **under 10 seconds**
  - Built and executed **27 Game Day** failure scenarios (routing, leader election, capacity tracking, peer communication); identified and fixed **5 critical edge cases** including session token priority mismatches, TTL recovery, and ALB cookie validation
- Operations & Full-CD:
  - Resolved **800+ tickets**; created **30+ composite alarms** reducing false positives by **70%**; reduced MTTR by **60%** through variant-level monitoring, priority fairness alarms, and cache error metrics
  - Enabled Full-CD for AppConfig, Canary, and Service pipelines with **20+ rollback alarms** and ECS circuit breakers; drove region expansion and **10+ model launches** (Claude Sonnets, Haikus, Opus, Amazon Nova family) resolving GMDS fallback, health check, and AZ blockers

Software Development Engineer II, Amazon – Q Business *July 2024 – Mar 2025*

- Accuracy Scorecard Framework:
  - Built Phase 2 evaluation pipeline across **5 ISV connectors** (SharePoint, WebCrawler, Confluence, GoogleDrive, S3) and **5 document types** (HTML, PDF, DOCX, PPTX, CSV) measuring Correctness (LLM-based), CitScore (citation validation), and Retrieval metrics (Recall/Precision @1/5/10) with automated CloudWatch emission and QuickSight dashboards
  - Implemented GoogleDrive test corpus (**115K** Finance FIQA documents, 300 queries) achieving **99.3% retrieval recall@1**; built SharePoint multi-document test sites with M365 token auth; created WebCrawler Amplify CDK infrastructure with HTML conversion scripts for multi-modal content (tabular, images, audio); achieved **60% reduction** in manual testing via automated canary infrastructure
- Data Ingestion Pipeline:
  - Implemented integration between ProductivityIngestionService and PreProcessor service handling UPSERT workflows, GetResource, and delete events; built IngestionNotificationLambda closing failure-scenario gaps and refactoring EventIngestionQueue model
  - Implemented dynamic attribute mapping for Kendra index ingestion via Q Business BatchPutDocument, simplifying complex data transformations for downstream indexing

Software Development Engineer, Amazon – AWS AppFabric *Nov 2021 – June 2024*

- Console Development & Launch:
  - Led end-to-end development of AWS AppFabric Console from **greenfield** project, contributing to **75%+** of Console ORR (Operational Readiness Review) tasks including pipeline setup, Cloudscape design system integration, localization, analytics, and security
  - Implemented operational dashboards (client-side and server-side), Content Security Policy violation testing using Puppeteer, and CSAT/Aperture feedback integration
  - Designed and built dynamic Announcements page improving customer communication about new product features and releases
  - Reduced console deep canary test runtime from **40 minutes to under 4 minutes** while maintaining full test coverage
  - Proposed and delivered console analytics strategy using Panorama for user behavior analysis, accessibility testing strategy, and Lighthouse security audits
- PostureHub API Design & Development:
  - Led API design and development for PostureHub across FrontEndService and Core Observability teams, implementing **6 APIs**: GetAppConfiguration, GetFinding, GetPosturePolicy, ListAppEntities, ListFindings, ListPosturePolicies
  - Proposed and implemented ARN structure for PostureHub resources ensuring scalability and maintainability
  - Collaborated with UX designer and Product Manager to resolve ambiguities in resource naming, schema design, and scope definition
- AppServer & Third-Party Integrations:
  - Co-delivered AppServer before re:Invent maintaining **90%+ test coverage**; built Fabric OAuth system ensuring message integrity for Asana, Slack, and G-Suite integrations
  - Re-invented Salesforce audit log strategy handling both hourly and daily log intervals, implementing kill switch for large EventLogFile sizes and event-level pagination design
  - Designed and implemented ListSheets API and Smartsheet strategy end-to-end, simplifying normalized schema and guiding performance testing
  - Created automated weekly data extraction ETL Lambda pipeline from DynamoDB (Asana/Slack) to CSV for Business Intelligence team
- IAM & Security:
  - Developed and released AWS AppFabric managed IAM policies (FullAccess & ReadOnly) end-to-end, coordinating cross-team reviews with AppSec, IAM, KMS, and documentation teams through MCM release process
  - Authored Service Description File (SDF) enumerating AppFabric actions, resources, and condition context keys for IAM integration
  - Onboarded Taj service for continuous API security testing, implementing automated security scanning workflows in CI/CD pipelines and production canaries
  - Participated as Guardian for threat model evaluations across EndUserPush, PostureHub, and Transmission services
- Cross-Team Leadership & Operations:
  - Mentored junior engineers, interns, and persistent team members; hosted office hours, conducted code reviews
  - Led JDK 8/11 to JDK 17 migration initiative within AppFabric, driving timeline discussions during Ops meetings
  - Contributed to correction of error(COE) preparation, cross-team pricing data alignment and onboarding documentation

Software Development Engineer, Amazon – Chime & WorkDocs *May 2021 – Nov 2021*

- Designed and implemented Partner Tag Customization feature for Amazon Chime, replacing the `<EXTERNAL>` tag with customizable `<PARTNER>` tags for trusted subsidiaries (NWCD), supporting multiple partners per conversation with feature flag rollback capability
- Proposed and demoed **3 rendering approaches**; implemented scalable solution using UCBuzzExpress SDK supporting Web and Desktop clients with multi-language translation support
- Led urgent migration of AWS WorkDocs Console from Angular to AWS Polaris (Cloudscape) design system within **3-month deadline**, driven by a critical security vulnerability in Angular; set up React testing framework with Babel configuration for KMS components
- Established greenfield project foundation including Polaris integration within Tangerine Box, internationalization support, and unit testing infrastructure; received appreciation from senior engineers for insisting on highest standards by proactively resolving UX discrepancies and driving clarity on product requirements beyond the original scope

Full Stack Engineer, Learning Equality *Feb 2021 – May 2021*

- Developed and maintained key features for Kolibri, an open-source educational platform serving **millions of learners** worldwide, using Django, SQL, and VueJS
- Led cross-functional project delivery coordinating frontend and backend milestones; optimized database queries and API endpoints improving system performance
- Contributed to code reviews, pair programming sessions, and agile development processes (sprint planning, daily standups, retrospectives)

Graduate Research Assistant, Old Dominion University *May 2019 – May 2020*

- Built responsive quiz application from scratch (student & admin views) used by multiple faculties and **100s of students** for English proficiency evaluation; built with HTML, CSS, JavaScript, PHP, and MySQL
- Developed parking space detection prototype for open parking lots using Python, OpenCV, Keras/TensorFlow, and CNN Sequential Model trained on ImageNet dataset achieving **90% accuracy**
- Improved existing real-time client portal UI efficiency by **70%** using MVC architecture, PHP, JavaScript, and MySQL
- Collaborated on Card Swiper Android application to read Student ID cards via RFID/NFC using audio jack device as part of campus-wide student services initiative

Graduate Teaching Assistant, Old Dominion University *Aug 2018 – May 2019*

- Trained undergraduate students in C++, Object-Oriented Programming, Research Strategies, and Information Literacy
- Conducted office hours and 1-on-1 sessions helping students improve overall academic performance

PROJECTS

- **Alpaca Trading Bot** Live Demo *Python, FastAPI, Next.js, Alpaca API, Docker*
  - Built an automated portfolio rebalancing bot that fetches stock rankings, compares week-over-week data, and executes trades across S&P 400/500/600 and NASDAQ 100 indices with configurable stock count and slack parameters; deployed with GitHub Actions scheduled rebalancing and Next.js dashboard on Vercel
- **Git-Ollama-Commit** *Shell, Ollama*
  - Built an Ollama-powered Git commit message generator that automates meaningful commit messages from staged changes; supports bash and zsh; 5 stars on GitHub
- **Build Your Own Claude Code (Rust)** *Rust, LLM, Tool Calling*
  - Built an LLM-powered coding assistant clone of Claude Code from scratch in Rust as part of CodeCrafters challenge; implemented OpenAI-compatible tool calling, agent loop, and multi-tool integration using HTTP RESTful APIs

EDUCATION

Old Dominion University, Master of Science in Computer Science 2018 – 2020
• Winner: HackU2019 Hackathon • Virginia Datathon – Certificate of Appreciation

Visvesvaraya Technological University, Bachelor of Engineering in Information Technology 2014 – 2018
• Best Outstanding Performer Award • National Level IT Fiesta'17 • TCS Tech Bytes