

Dinesh Bali

✉ dineshbali45@gmail.com ☎ +1-660-528-8938 in LinkedIn 🌐 github.com/dineshbali18

SUMMARY

Software Engineer with nearly four years of experience building distributed systems, cloud-native platforms, and AI infrastructure. Proven track record of developing scalable backend services, multi-agent AI systems, and event-driven architectures for production environments.

TECHNICAL SKILLS

- Backend:** Golang (Echo), Python (FastAPI), Java (Spring Boot, Spring MVC, Spring Security, Spring Data JPA), Node.js (Express)
- Distributed Systems:** Event-Driven Architecture, Microservices, Kafka, Redis Streams, Redis Pub/Sub, WebSockets, gRPC, REST APIs, Webhooks, Concurrency, Idempotency
- Databases:** PostgreSQL, MySQL, Redis, DynamoDB, Pinecone
- Cloud & DevOps:** AWS (EC2, VPC, ALB, S3, RDS, SQS, ElastiCache), Docker, Terraform
- Frontend:** React, TypeScript, JavaScript, Redux, Next.js
- Machine Learning & Deep Learning:** TensorFlow, Scikit-learn, Transformers, Transfer Learning, Model Evaluation
- Generative AI:** Multi-Agent Systems, RAG, AI Orchestration, Agent Evaluation, Model Context Protocol (MCP), Vertex AI, Langfuse, Hugging Face, PyTorch, LLM Fine-Tuning (LoRA, QLoRA, PEFT), Supervised Fine-Tuning (SFT)
- Developer & AI Tools:** Git, Postman, Swagger, Prometheus, Grafana, Claude, Codex
- Networking & Systems:** Linux, TCP/IP, HTTP/HTTPS, DNS, TLS/SSL, Load Balancing, Reverse Proxies

EXPERIENCE

Founding Software Engineer Jul 2025 – Present
Alpheva AI - Fintech startup New York, USA

- Architected a multi-agent AI platform on Vertex AI using Google ADK, enabling adaptive model routing, parallel tool orchestration, and grounded web search, expanding financial insight coverage by 75%.
- Built a production-scale RAG knowledge ingestion platform using PyMuPDF, S3, DynamoDB, and Pinecone, enabling asynchronous document processing, semantic indexing, and low-latency retrieval, reducing retrieval latency by 65%.
- Engineered an evaluation platform using Langfuse and domain-specific financial benchmarks, automating RAG and agent validation across 500+ benchmark scenarios to prevent production regressions.
- Scaled Redis Streams processing with sharded consumers, concurrent workers, retries, and DLQs, increasing throughput by 4× while processing 100K+ events/day.
- Implemented idempotent Go webhook ingestion for Plaid, Stripe, and Array with deduplication, batched upserts, and transactional writes, reducing database load by 60%.
- Provisioned highly available AWS infrastructure across VPC, EC2, ALB, RDS, SQS, S3, and ElastiCache, supporting distributed backend microservices with 99.9%+ service availability.

Graduate Assistant Software Developer May 2024 – May 2025
Northwest Missouri State University Missouri, USA

- Owned the end-to-end backend platform for 6+ production applications in Go and Java (Spring Boot), leading a team of 8 engineers and establishing scalable, fault-tolerant service patterns that enabled reliable weekly releases.
- Designed Kafka and Redis-backed messaging patterns for Spring Boot services, improving reliability across real-time and asynchronous workflows.
- Migrated chat persistence to Kafka-backed workers, reducing latency by 40% and increasing write throughput by 60%.
- Implemented Redis Pub/Sub fanout for WebSocket chat, supporting 5K+ concurrent connections and 10K+ messages/min.
- Built Prometheus and Grafana monitoring for Dockerized services, cutting detection time by 50% and resolution time by 30%.

Software Engineer Feb 2023 – Aug 2023
Bobble AI - Series C Haryana, India

- Engineered high-throughput Go microservices and a sharded Redis Streams pipeline with consumer groups, supporting 50M+ MAU and high-volume event processing.
- Developed a TensorFlow inference service for real-time intent classification, enabling low-latency predictions for keyboard applications.
- Designed resilient event-driven services with retries, backoff, and load shedding, reducing system errors by 30%.
- Optimized MySQL performance by tuning indexes, query plans, connection pooling, and caching, reducing P95 latency by 45%.

Software Engineer Jan 2022 – Jan 2023
Cognizant Telangana, India

- Implemented scalable Spring Boot microservices on AWS using concurrency and multithreading, reducing API latency by 40%.
- Developed RESTful APIs with Spring MVC, Spring Security, and Spring Data JPA, supporting authentication, authorization, and persistence workflows.
- Improved React and Redux frontend workflows through centralized state management, increasing UI responsiveness by 20%.

EDUCATION

Northwest Missouri State University Jan 2024 – May 2025
Master of Science in Computer Science GPA: 4.0/4.0

GITAM University Jun 2018 – Apr 2022
Bachelor of Technology - Computer Science GPA: 8.37/10.00