# Avinash Kumar

✉ k8.avinash@gmail.com ☎ +91 7678520633 Bangalore linkedin.com/in/avinash09 leetcode.com/avinashh

## ABOUT

A Software Engineer with 3+ years of experience in backend technologies in GenAI, Infra and Search domain.

## PROFESSIONAL EXPERIENCE

**Samsung R&D Institute**

*Senior Software Engineer* Mar 2023 – present | Bengaluru, India

- Working as part of **Search Team,** developed search system for Samsung website/app results **boosting sales up to 3 times**. Facilitating search over **30+ countries.** expanding it to **80+ countries.**
- Business coordination for boosting results of product's latest launch/offer as per latest product release.
- Worked as part of **Language AI Team. Developed QNA (RAG) system** for **Bixby 3.0** feature for Device QNA. allowing users to talk to devices with personalized context (Rolled out in Samsung Fold 6) using microservice architecture.

.

- Load & Stress testing on LLM inference server e.g. TGI, TensorRT LLM measuring latency/accuracy tradeoff with/without multiple capabilities (page, flash attention) with/without quantization in multithreaded/concurrent environment for optimal performance on various models (Llama, Mistral, Samsung LLM) using locust and python.
- LLM Model conversion toolkit from Open Neural Network Exchange (ONNX)/TRTengine format supported by tensorRT LLM (bash, docker, ECS)
- Developed Prompt Studio, a workspace and collaboration platform, that allows prompt developers to create/test prompts with centralized versioning control system for bixby3 prompt templating.
- Led the development of **Large Language Model (LLM)-powered chatbot** implementing context-based **Retrieval-Augmented Generation (RAG) System** enabling query in NLG.

**Tech Stack**: Java, Python, Spring Boot, FastAPI, Microservices, Flask, Node.js, Reactjs, MySQL, PostgreSQL, Redis, Rabbit MQ, Rest API, Opensearch, AWS, EKS.

*Software Engineer* Aug 2022 – Mar 2023 | Bangalore, India

- **Infrastructure migration** to Kubernetes (EKS), enhancing availability and minimizing downtime. **Set up infrastructure** for **centralized** logging using the **ELK stack**. **Cross-region code unification** enhancing its maintainability. ( Kubernetes, Docker, AWS, Github Actions, EKS).
- Added media support along with text for Samsung Chatbot-IN (Whatsapp) where curator can upload and map media with specific questions.

**Tech Stack:** Python, Flask, Javascript, Microservices, EKS, Docker, Github Actions, MySQL, Redis.

**Cognizant Infrastructure Services,** *Programmer Analyst* Jul 2021 – Aug 2022 | Kolkata, India

- Agent service portal - Developed user ticket lifecycle management for the Agent service portal.
- Supported production bug fixes (hotfix) for same. Followed agile methodology to deliver results.

**Tech stack** - Javascript, Angular, ServiceNow

## EDUCATION

**Bachelor Of Technology,** *Lakshmi Narain College of Technology & Excellence* Aug 2017 – Jun 2021 | Bhopal, India

CGPA: 8.34

## SKILLS

**Programming Languages:** Java, Python, Javascript, SQL

**Libraries / Frameworks:** Flask, Spring Boot, Node.js, React.js, Locust

**Tools / Platforms:** AWS, Elasticsearch/Opensearch, Kibana, Redis, Rabbit MQ, Docker, Kubernetes, Nginx, Github Actions, Argo CD.

**Databases:** MySQL, PostgreSQL, Redis, Elasticsearch.

## CERTIFICATES

- AWS Networking
- Docker and Kubernetes - The Complete Guide

## RECOGNITION

**Samsung Excellence Award**

Star of the Quarter Award(Q3) presented for Extraordinary effort in the Gen-AI project