

SARATH KUMAR

Bengaluru, Karnataka, India

+91 90808 07001

sarathofficial11@gmail.com

linkedin.com/in/sarath-kumar-3b621316b

workwithsarathkumar-hash.github.io

Forward Deployed Engineer — Agentic AI · RAG · LLM
Document Intelligence in Production

Forward Deployed Engineer, **7+ years** shipping production systems embedded with enterprise client teams from discovery to operational handover. Leads the **Generative AI team at Mphasis: LLM + OCR/ICR document intelligence** and decision-automation pipelines in regulated financial environments, owned end-to-end. Builds **agentic workflows** — RAG, tool calling, multi-step orchestration, eval and quality gates — on **AWS** with Python, Node.js and Java, and frames accuracy / latency / cost trade-offs in decision-ready language for client executives.

CORE COMPETENCIES

Agentic AI / LLM	RAG · tool & function calling · multi-step orchestration · prompt engineering · evals & quality gates · OCR/ICR document intelligence · vector databases
Delivery	Customer-embedded discovery → production · architecture ownership · POC-to-production adoption · executive stakeholder influence · mentoring
Backend & Cloud	Python · Node.js · Java · REST/SOAP APIs · AWS Lambda services · data integrations · CI/CD and environments · load testing · deep debugging
Security & Reliability	PII handling · data tokenization (Protegrity) · access controls · retries, fallbacks and safe failure modes · monitoring and observability
Conversational AI	Kore.ai · Amazon Connect · Amazon Lex v2 · Google CCAI / Dialogflow · virtual agents · contact-center modernization

EXPERIENCE

Senior Software Engineer — Forward Deployed Engineer, Generative AI Team Lead

Feb 2025 – Present

Mphasis · Bengaluru, India

AI: enterprise LLM + OCR/ICR document intelligence and decision automation for financial services

- Lead the **Generative AI** team as an embedded FDE with financial-services clients — scoping ambiguous problems with their stakeholders and shipping enterprise-scale AI systems into their own stack.
- Own architecture and technical roadmap for **LLM extraction, validation and summarization pipelines** over high volumes of complex unstructured documents, discovery through rollout and operational readiness.
- Engineer for enterprise-grade quality in a **regulated environment**: accuracy and confidence thresholds, human-review fallbacks, scalability and production-readiness across data → validation → business-action flows.
- Frame **accuracy / latency / cost** trade-offs for clients and cross-functional partners in decision-ready terms; mentor engineers on reusable AI delivery patterns.

Senior Software Development Engineer

May 2024 – Jan 2025

MResult Tech · Bengaluru, India

AI: client-facing RAG applications, end-to-end from discovery through deployment

- Worked directly with the client end-to-end — discovery and requirements through design, development, and production deployment — as the embedded engineer on their Generative AI initiative.
- Built **LLM applications** and cloud services pairing conversational interfaces with Generative AI over customer data stores.
- Implemented **Retrieval-Augmented Generation** to ground models in live enterprise data, raising answer quality and decision usefulness.
- Owned ingestion into **vector databases** — chunking, indexing, refresh — tuned for retrieval scalability, speed and accuracy.

Conversational AI Developer — Generative AI

Mar 2024 – May 2024

AI: Generative AI in the live response path of customer journeys

- Designed and delivered **Kore.ai** conversational solutions with Generative AI generating responses for production customer journeys.
- Worked cross-functionally to tune agent performance and streamline the operational processes around it.

Management Consulting Analyst — Cloud Contact Center Delivery

Oct 2021 – Dec 2023

Accenture · Bengaluru, India

AI: client-embedded conversational AI (Lex v2) delivery for a US bank's production contact center

- Embedded directly with a US banking client's engineering and operations teams, leading their contact-center transformation onto a cloud-native **Amazon Connect** architecture end-to-end.
- Engineered complex **Lex v2 Lambda** services and multiple conversational surfaces (chat, voice, virtual agents); resolved intent ambiguity via a PutBotSession proof-of-concept carried into the production design.
- Integrated conversational-AI Lambdas with **Protegrity tokenization** for secure handling of sensitive customer data.
- Built an IVR journey reporting system (VistaString) for measurable visibility into containment; designed a load-testing tool simulating **1,000+ concurrent users**, and a SOAP API delivered across Python, Node.js and Java.

Associate Engineer — Technology

Jul 2019 – Oct 2021

Virtusa Consulting · Chennai, India

AI: production chatbots and Dialogflow / CCAI virtual agents for US and APAC clients

- Launched an **AI chatbot** for a US health-insurance client, improving issue resolution and reducing ITSM team load.
- Built a **Google Dialogflow / CCAI** virtual agent for an Australian grocery retailer, streamlining product access for customers.
- Migrated a US auto-insurance client's contact center to Amazon Connect (Holiday Management, Voicemail); drove Center-of-Excellence R&D integrating channels across Facebook, WebEx, Telegram and WhatsApp.

SELECTED PERSONAL PROJECTS

MAX — Personal AI Operating System

2026 – Ongoing

AI: multi-agent orchestration, tiered local/cloud reasoning

- Architected a **multi-agent system** — specialist agents for website, GitHub, LinkedIn, portfolio strategy, advisory, and PR — orchestrated through a tiered local/cloud LLM router.
- Built a **risk-classifying safety and confirmation layer** gating every tool call; integrated a voice front-end, home-automation control, and real tool-calling across the stack.

Mental Health SLM (Prototype)

2026

AI: SLM fine-tuning, quantization, applied AI safety

- Fine-tuned a 1.5B-parameter model via **LoRA** with **4-bit quantization** on Apple Silicon (MLX), trained on a synthetic dataset built from cognitive-behavioral-therapy methodology.
- Designed a **deterministic crisis-safety gate** that structurally excludes risk content from the model rather than relying on model judgment — advisory-only by architecture, with documented scope limits.

ResumeSignal

2025

AI: privacy-first LLM application

- Built and shipped a **privacy-first resume scorer**, live in production — nothing stored, nothing collected.

EDUCATION

B.E. Computer Science & Engineering

Knowledge Institute of Technology (KIOT), Salem, India · 2015–2019

CERTIFICATIONS

Solution Training for Amazon Connect · Data Science Methodology · Data Analysis with Python · Open Source Tools for Data Science · Migrating Databases & VMs to Google Cloud Platform

Personal website, including its intro video, built and deployed end-to-end with an **agentic AI workflow**. Languages: English (professional working).