

BAIBHAV KUMAR | baibhav1kumar@gmail.com | +91 9315195961 | Bangalore
<https://linkedin.com/in/baibhav-kumar-> | <https://www.github.com/baibhavKumar1>

SUMMARY

Full Stack Generative AI Engineer with 2+ years of experience designing, building, and deploying production-grade AI applications. Expertise in Retrieval-Augmented Generation (RAG), multi-agent AI systems, prompt engineering, embedding pipelines, vector databases, and LLM fine-tuning for domain-specific generative models. Hands-on experience with LangChain, LangGraph, LlamaIndex, Google ADK, Hugging Face, Vertex AI, Docker, and cloud-native deployments across AWS and GCP. Skilled in model evaluation, responsible AI, production monitoring, and scalable REST API development.

PROFESSIONAL EXPERIENCE

Full Stack Developer / Mavoix Technology Solutions Pvt. Ltd. | June 2024 - Present

- Building AI-enabled solutions for **Aditya Birla Group**, a conglomerate operating across countries
- Developed enterprise-grade AI applications using Google ADK, AG-UI, and Gemini API on GCP
- Deployed scalable AI solutions using Cloud Run and Vertex AI
- Designed and implemented enterprise Retrieval-Augmented Generation (RAG) pipelines using **LangChain, LangGraph, and LlamaIndex**, building OCR-driven ingestion pipelines that transformed enterprise documents into structured JSON-LD, optimized chunking, embedding generation, top-k retrieval, reranking, and vector search for high-accuracy knowledge retrieval.

Full-Stack Development

- Built and maintained end-to-end web applications using React, Next.js, and Node.js (1) (2)
- Developed RESTful APIs and microservices for enterprise clients, handling 10K+ requests per day
- Implemented frontend components and responsive UI with React/Next.js and Tailwind CSS
- Deployed applications on AWS (EC2, S3, Lambda) cloud infrastructure

AI/ML Development

- Built LLM-based chatbots with custom embeddings and vector database integration, improving customer query resolution by 40%
- Implemented Google ADK-based agents with tool use, memory, and multimodal reasoning
- Set up Cloud and DevOps CI/CD pipelines for ML and web applications using GitHub Actions

KEY PROJECTS

Wishop (Whatsapp Automation System) - wishop.xyz

- Usable with whatsapp business app, helping you to continue your conversation from anywhere.
- Able to linked with google sheets (public & private) to manage the automations
- 15 paid users with 2 mins of onboarding time.

MusicBuddy — AI Music Generation Agent - musicbuddy.ai

- Built a conversational AI agent orchestrating multi-step workflows for text-to-music generation with intelligent tool selection.
- Fine-tuned open-source music generation models on domain-specific datasets and optimized prompt engineering strategies including zero-shot, few-shot, and iterative prompting.
- Designed prompt versioning and rubric-based evaluation pipelines to improve generation quality and query coverage.
- Built Langfuse observability and implemented responsible AI guardrails to prevent copyrighted music generation.

- Clouderly (Agent Management System)** - github.com/baibhavKumar1/tinycloud
- Engineered & Deployed High-Throughput API: Designed and deployed a high-performance Agent API microservice on Google Cloud Platform (GCP), handling 100+ daily AI agent requests.
 - Cloud Architecture & Cost Optimization: Built serverless infrastructure using GCP (Cloud Run, IAM, Secret Manager), reducing infrastructure operational costs by 35% while maintaining 99.9% service uptime.
 - Workflow Automation & Agent Execution: Integrated dynamic AI tool execution and vector context retrieval pipelines, boosting automated task execution throughput by 45%.
 - Security & Access Control: Implemented strict OAuth2 authentication, rate limiting, and fine-grained GCP IAM roles, eliminating unauthorized endpoint access and reducing API abuse by 99%.

EDUCATION

BTech in Computer Science & Engg.	2023
Masai School	2024