

Sameer Bagul

AI Engineer | Software Engineer | Full Stack | LLMs | RAG

Pune, India | +91 7841941033 | sameerbagul2004@gmail.com

linkedin.com/in/sameer-bagul | github.com/Sameer-Bagul | sameerbagul.com

Summary

Software Engineer and AI Engineer with experience building production-ready AI and full-stack applications, backend services, REST APIs, databases, automation workflows, and cloud deployments. Strong in Python, TypeScript, JavaScript, Node.js, React, SQL, Docker, CI/CD, and software architecture, with hands-on experience building scalable APIs, distributed workflows, testing automation, and production systems. Specialized in LLM applications, RAG, agentic AI, semantic search, and integrating AI capabilities into customer-facing products.

Experience

Revccloud

AI Engineer

Pune, India

Aug. 2026 – Present

- Engineer production-ready AI and software capabilities across Revcloud's B2B lead-generation and revenue funnel, owning features from PRD and system architecture through implementation, testing, deployment, and monitoring.
- Develop AI-driven traffic and lead-routing systems that combine semantic signals, funnel analytics, and revenue outcomes to improve traffic allocation and routing decisions.
- Contribute to **Alyson.ai**, an AI-powered customer-acquisition platform for generating and optimizing landing pages, routing traffic, capturing leads, and automating customer interactions.
- Build LLM applications and agentic workflows using multiple model providers, LangChain, LangGraph, RAG, hybrid retrieval, semantic search, and production APIs across Node.js, Python, React, Next.js, and Astro.
- Develop an AI meeting-intelligence workflow that processes conversations through transcription, chunking, semantic retrieval, and agentic post-processing to generate structured notes, action items, and automated organizational follow-ups
- Build and operate AI systems on AWS with Docker, CI/CD, logging, monitoring, evaluation, and LLMOps practices across the development-to-production lifecycle.

Dimakh Consultants

AI Developer

Pune, India

Apr. 2026 – Aug. 2026

- Built and deployed production software and AI services for business workflows across sales, customer support, and lead engagement, owning features from requirements and system design through implementation, API integration, testing, and deployment.
- Integrated an AI assistant into **Sudapedia**, a production business application for Sudarshan Chemical Industries, bringing LLM-powered retrieval and domain-specific assistance directly into the sales workflow.
- Built an AI customer assistant for a client website, combining LLM orchestration, business knowledge, retrieval, and application APIs to handle customer questions and support pre-sales interactions.
- Developed AI calling workflows for sales teams, working on conversation orchestration, prompt design, token optimization, and backend integrations to improve inference efficiency without degrading conversation quality.
- Designed RAG and agentic workflows that connected business knowledge, application data, and external APIs to produce context-aware responses for customer-facing AI systems.

Bug0

Software Engineer

Remote

Nov. 2025 – Mar. 2026

- Built core components of Bug0's AI-native testing platform, developing agentic browser-testing workflows around Playwright for generating, executing, and maintaining end-to-end regression tests.
- Developed Node.js backend services and CI/CD integrations using Docker and GitHub Actions, reducing deployment cycle time by **70%**.
- Worked directly with customer engineering teams, including **Vercel** and **Drumwave**, to integrate testing workflows, debug production issues, and adapt the platform to real application environments.
- Built supporting tooling around test execution, failure analysis, and AI inference to improve reliability and reduce friction in automated QA workflows.

LabsCheck

Full-Stack Developer & DevOps Engineer Intern

Pune, India

Aug. 2025 – Oct. 2025

- Built and shipped features for LabsCheck, a healthcare marketplace for searching, comparing, and booking diagnostic tests across **12,000+** verified pathology labs.
- Developed backend APIs and application services across Next.js and Python, supporting lab discovery, test-price comparison, booking workflows, and integration with external services.
- Built an AI-powered report-generation workflow as a parallel product initiative, designing the APIs and service integration required to turn report data into structured, user-facing outputs.
- Owned deployment and infrastructure for multiple services, configuring Docker, Nginx, and CI/CD on Contabo VMs and other environments and handling production releases, debugging, and service reliability.
- Automated lead and marketing workflows with AI agents, supporting **100,000+ leads** for outreach and follow-up operations.

Selected Software & AI Projects

NoteForge AI | *RAG, LangGraph, LLM Orchestration, Vector Search*

GitHub

- Built a full-stack AI-powered YouTube-to-Notes platform using React 18, TypeScript, Vite, Tailwind CSS, Node.js, and Express, converting videos/playlists into structured study notebooks through an automated 5-stage processing pipeline.
- Engineered a hybrid LLM orchestration layer with Ollama, LM Studio, Google Gemini, and xAI Grok, including provider routing, task-based model mapping, fallback execution, JSON repair, and token-stream generation for resilient AI inference.
- Implemented advanced RAG and agentic workflows with LangChain + LangGraph, combining semantic chunking, BM25 + vector ensemble retrieval, hallucination self-critique loops, optional Pinecone integration, real-time SSE job streaming, and AI-generated mind maps, quizzes, and interview Q&A.

Athena | *Agentic AI, LangGraph, Voice, RAG, MCP*

GitHub

- Built a production-grade Electron.js + TypeScript + React/Vite desktop AI assistant with a real-time Three.js/VRM 3D avatar engine, delivering immersive human-like interactions at approximately **60 FPS**.
- Architected a modular agentic AI system using LangGraph/LangChain with multi-provider routing across Google Gemini, Ollama, LM Studio, and OpenAI-compatible APIs, enabling autonomous tool use, multi-step reasoning, and dynamic model orchestration.
- Implemented a low-latency local voice pipeline with Python Whisper STT and Node SuperTonic TTS, alongside hybrid memory, local RAG/vector retrieval, MediaPipe facial tracking, lip-sync blendshapes, and MCP tool connectors, reducing idle CPU overhead by approximately **66%**.

JalSakhi | *ML, LLMs, FastAPI, LangChain, MLOps*

GitHub

- Built an end-to-end AI-powered precision agriculture platform using React Native (Expo + TypeScript), Node.js/Express, MongoDB Atlas, and FastAPI for irrigation planning, analytics, and multilingual farmer/admin workflows.
- Designed and productionized MLOps pipelines for **3 ML services + 1 LLM service**, covering crop-water prediction, soil-moisture forecasting, village water allocation, and a Groq Llama-3.1-8B + LangChain chatbot with tool-calling across ML endpoints.
- Trained models on datasets ranging from **2,873 to 118,013 engineered samples** with reproducible validation splits, and implemented secure API controls including API-key authentication, 100 req/min rate limiting, 60s timeouts, 1MB payload guards, and cross-device distributed testing.

Technical Skills

AI & Generative AI: LLMs, Generative AI, RAG, Hybrid RAG, Agentic AI, AI Agents, Prompt Engineering, Semantic Search, Embeddings, Vector Search, LLM Evaluation, Tool Calling

AI Frameworks & Models: LangChain, LangGraph, LlamaIndex, Hugging Face, Ollama, Google Gemini, OpenAI-compatible APIs, Groq, PyTorch

Programming: Python, TypeScript, JavaScript, Java, C, C++, SQL

Backend & APIs: Node.js, Express.js, FastAPI, Flask, REST APIs, Microservices, OpenAPI, SSE

Data & Databases: PostgreSQL, MongoDB, Redis, FAISS, Pinecone, BM25, Vector Search, ETL, Data Processing

Cloud, DevOps & Infrastructure: AWS, Docker, GitHub Actions, CI/CD, Nginx, Cloudflare, Linux, Contabo

Observability & Engineering: Prometheus, Grafana, Structured Logging, Monitoring, API Testing, Git

Achievements & Technical Leadership

- **Smart India Hackathon 2024 — National Runner-up:** Built an AI-powered adaptive skills-learning system using TypeScript, Node.js, and React, with an agentic learning workflow that dynamically updated user roadmaps from weekly assessments; built data-scraping and semantic-search pipelines to prepare relevant knowledge for AI-driven learning.
- **Innovate You Hackathon 2025 — National Winner, AI/Automation:** Built an end-to-end AI-powered process-automation platform focused on using agentic workflows to automate multi-step tasks and reduce manual operational effort.
- **India Innovates 2026 — National Finalist:** Selected among the top **5,000+ teams** from more than **1.26 crore applicants** with JalSakhi, an AI-powered precision agriculture platform combining **3 ML services + 1 LLM service**, LangChain tool calling, and API-driven irrigation and water-allocation workflows.
- **Technical Lead & Technology Advisor, APP Club:** Mentored **200+ engineering students** through AI and full-stack projects, system design workshops, and technical sessions; advised student teams on architecture, implementation strategy, and engineering best practices.

Education

PES Modern College of Engineering

Bachelor of Technology in Information Technology

Pune, India

2022 – 2026

- CGPA: **7.77/10** | MHT-CET Percentile: **94.95**
- Relevant coursework: Data Structures & Algorithms, Machine Learning, Database Management Systems, Operating Systems, Software Engineering, Distributed Systems.