

# Prathamesh Saraf

Senior Forward Deployed Engineer | GenAI Architect | Tech Lead  
US (Remote) • pratamesh1867@gmail.com • psaraf.pages.dev

## EDUCATION

### INDIAN INSTITUTE OF SCIENCE

M.TECH (RES) IN COMPUTATIONAL  
AND DATA SCIENCE

Apr 2024 | Bangalore, India

CGPA: 8.1 / 10.0

### DKTES' TEXTILE AND ENGINEERING INSTITUTE

B.TECH IN COMPUTER SCIENCE

May 2020 | Kolhapur, India

Magna Cum Laude

CGPA: 9.24 / 10.0

## LINKS

Github:// [S1LV3RJ1NX](#)

LinkedIn:// [sarafpr](#)

Site:// [psaraf.pages.dev](#)

Book:// [adventures-with-llms](#)

## SKILLS

### LANGUAGES

Python • TypeScript • C++ • SQL

### GENAI / LLM

LangGraph • LangChain • LlamaIndex •  
Google ADK • LiveKit • LiteLLM • MCP  
• AI Gateway • vLLM • Langfuse

### BACKEND / INFRA

FastAPI • Postgres + pgvector • Qdrant •  
OpenSearch • Redis • Docker •  
Kubernetes • AWS • GCP • Cloudflare

## CONTRIBUTIONS

**Speaker** – LF MCP Dev Summit,  
Bengaluru (Jun 2026).

**Toptal** – Top 3% Talent, AI Specialization.

**Author** – *My Adventures with LLMs*  
(Leanpub).

**TrueMem**: *A Model-Agnostic Memory Layer  
for AI* – TrueFoundry, 2026 (byline).

**Scalable and Cost-Effective Voice Agents** –  
CVS Health Tech Blog, 2025.

**CARL**: *Cost-Optimized Online Container  
Placement on VMs using Adversarial RL* –  
IEEE Trans. on Cloud Computing, 2025.

## EXPERIENCE

### TRUEFOUNDRY | SR. FORWARD DEPLOYED ENGINEER

Feb 2024 -- Present | USA (Remote)

- Leading the GenAI platform build-out for **CVS Health** – voice agents, agentic workflows, RAG, LLM inference and fine-tuning – serving **hundreds of millions of member calls**.
- Architected the production voice-agent platform on **LiveKit + LangGraph + LiteLLM** replacing legacy IVR; drove **~95%** containment on outbound flows and cut voice infrastructure spend **~50%**. Documented in the CVS Health Tech Blog.
- Designed **TrueMem** – a model-agnostic two-tier memory service (~10 ms similarity, ~45 ms context prep, semantic dedup, importance-scored lifecycle); signed byline on the TrueFoundry engineering blog.
- Shipped TrueFoundry's public **MCP Gateway Registry** – a catalog of MCP servers with OAuth flows across heterogeneous identity providers (**Google, GitHub, Microsoft, Okta**) and scope-aware tenant identity.
- Forward-deployed solutions architect across enterprise accounts including **Merck, Otsuka, Siemens**; major contributor to **Cognita (4.4k+ stars)**, an OSS production RAG framework.

### CHATOWL | TECHNICAL LEAD | FOUNDING ENGINEER

Dec 2022 -- Jan 2024 | USA (Remote)

- Owned the technical roadmap and full product stack for ChatOwl's AI therapeutic-sessions platform; shipped successive releases at a Series-A startup and partnered with founders on product strategy.
- Led design across Conversational AI, backend, LLM fine-tuning, datastores and DevOps; mentored the engineering team and established release / on-call practices.

### SAARTHI.AI | CHATBOT DEVELOPER

Aug 2020 -- Aug 2021 | Bangalore, India

- Designed multilingual **text and IVR**-based chatbots; extended RASA's open-source core with automated conversation testing, improving **testing effort by 50%**.
- Stood up analytics services that fed business leads back into product (**grew customer outreach 20%**); containerized on serverless infrastructure, **cutting cloud cost by 15%** versus VM-based deployment.

## PROJECTS

- mcp-guardian** – MCP proxy that replaces hundreds of tool schemas with three meta-tools; **99.7%** startup-token reduction (160k → 456), OAuth-aware fan-out, per-scope allowlists. Featured talk at LF MCP Dev Summit (S1LV3RJ1NX/mcp-guardian).
- Yukti** – Open-sourced multi-tenant Retrieval-Augmented Generation platform: FastAPI + ARQ workers, Unstructured.io / Docling parsers, pgvector + Qdrant hybrid search, LiteLLM gateway, Vite/React UI (S1LV3RJ1NX/yukti).
- AIME** – Open-sourced meeting-intelligence and voice-agent platform: LiveKit Agents, LangGraph orchestration, Google Calendar / Meet integration, FastAPI, Postgres (S1LV3RJ1NX/AIME-backend).