

SHREY KADIA

Senior AI Engineer — Agentic Systems, Real-Time Voice AI, RAG & LLM Evaluation

srkkadia@gmail.com | github.com/srkkadia | linkedin.com/in/srkkadia

PROFESSIONAL SUMMARY

Senior AI Engineer with 3.5 years shipping production LLM systems end-to-end. Architected a **multi-agent platform on AWS Bedrock** spanning 6 product apps with **28 domain agents**, **~640 tools** and a **~640-entry RBAC permission layer**. Built two production voice AI systems: an **AI voice interviewer (LiveKit)** across **5 languages**, and a self-hosted **AI calling agent (FreeSWITCH/SIP)** with **~700 ms** response latency. Also fine-tuned **Qwen2.5-3B (QLoRA)** with a custom evaluation harness, and earlier delivered ML risk models that improved accuracy **15%**.

TECHNICAL SKILLS

LLM & Agentic Stack: AWS Bedrock (Claude, Converse API), Multi-Agent Orchestration (ReAct pattern), Tool/Function Calling, Prompt Caching, LLM Fine-Tuning (QLoRA, Unsloth), HuggingFace Transformers

Retrieval & Data: RAG, pgvector, AWS Titan Text Embeddings V2, Semantic Search, PostgreSQL, SQL, TypeORM, Multi-Tenant (Schema-per-Tenant) Architecture

Evaluation & Safety: Hallucination / Groundedness Detection, JSON & Schema-Validity Scoring, Per-Field Exact-Match, Jaccard Similarity, RBAC & Guardrail Design

Voice AI & Real-Time Systems: LiveKit Agents SDK, FreeSWITCH / SIP Telephony, Deepgram (STT/TTS), Sarvam (STT/TTS), Silero VAD, Real-Time Turn-Taking, WebRTC, Server-Sent Events (SSE)

MLOps & Cloud Infra: AWS (Bedrock, Lambda, Step Functions, S3, EC2, ECS, Athena), Docker, Redis, BullMQ, MLflow, GitHub Actions, Zod / Pydantic Schema Validation

Languages & Frameworks: Python, TypeScript/Node.js (NestJS), FastAPI, LightGBM, XGBoost

PROFESSIONAL EXPERIENCE

Senior AI Engineer | Mehta Softech Pvt. Ltd., Ahmedabad

Sept 2025 – Present

- **Agentic platform architecture:** architected a multi-agent AI system (AWS Bedrock + custom ReAct-pattern orchestrator in NestJS) spanning **6 product apps** with **28 domain agents** and **~640 tools**, replacing manual multi-click workflows with natural-language task completion that reduced average recruiter task time from **~1.5 min to ~15 sec** for common workflows.
- **Tool & guardrail design:** built the agent tool layer (Zod schemas) and a declarative **RBAC permission layer (~640 entries)** that hides forbidden tools before the LLM sees them, enforcing per-action authorization across a **multi-tenant, schema-per-tenant** Postgres deployment.
- **RAG / semantic search:** implemented natural-language candidate search using AWS Titan Text Embeddings V2 (1024-dim) + pgvector cosine similarity, with a **24-hour** query-embedding cache cutting redundant Bedrock calls.
- **AI calling agent (CRM):** engineered a self-hosted AI phone-agent platform (**FreeSWITCH/SIP**, NestJS, AWS Bedrock Claude, Sarvam STT/TTS, Silero VAD) that **auto-calls every new lead within 15 min**, qualifies it (**HQL/MQL/LQL**), books callbacks and meetings, and writes the call log, recording and transcript back to the CRM.
- **Real-time turn-taking engine:** built a custom conversation engine (VAD turn-hold, preemptive LLM generation, barge-in handling, sentence-streamed TTS) that reaches **~700 ms end-to-end response latency** on live phone calls.
- **Scalable voice infrastructure:** designed a **stateless, horizontally scalable** stack that handles **20 concurrent calls per server**, with a queue-based dialer that respects calling hours, retries on busy or unanswered calls, and bills per answered minute.
- **AI voice interviewer (ATS):** shipped a production AI voice interviewer (**LiveKit Agents SDK**, **Deepgram/Sarvam TTS**, **Silero VAD**, **AWS Bedrock**) conducting live, resume- and JD-adaptive screening interviews across **5 languages** and **3 interview modes**; self-hosting-ready, in production on LiveKit Cloud. It has conducted **50+ screening interviews** saving an estimated **25+ hours**.
- **Serverless evaluation pipeline:** designed an AWS Step Functions workflow (**8 chained Lambdas**) automating transcription, multi-face/voice integrity checks, abusive-language screening, and LLM-based interview scoring into a single automated report, removing manual reviewer effort from first-round screening.
- **LLM fine-tuning + eval harness:** built a data pipeline (deterministic extraction → LLM labeling → hallucination/groundedness auditing → stratified splitting) and fine-tuned **Qwen2.5-3B** via QLoRA (Unsloth); built an evaluation harness scoring JSON/schema validity, per-field exact-match, and Jaccard skill-overlap, then deployed locally via GGUF/Ollama for private, low-cost inference.

Data Scientist | Grownited Pvt. Ltd., Ahmedabad

Apr 2023 – Sept 2025

- Independently built and deployed loan-risk scoring models, improving accuracy by **15%** and cutting loan-approval turnaround from 48 to **36 hours** (25% faster).
- Engineered **30+** features from transactional and bureau data, identifying 5 key variables that reduced false positives by **18%**.
- Owned the full ML lifecycle — training (LightGBM, ANN), experiment tracking (MLflow), and production deployment (FastAPI on AWS EC2) — cutting deployment time by **40%**.
- Defined 6 core credit-risk KPIs that directly informed risk-policy updates and improved portfolio segmentation accuracy by **22%**.
- Optimized **15+** SQL workflows, reducing query execution time by **20–40%** and dashboard latency from 5 minutes to under **3**.

EDUCATION

Bachelor of Engineering (B.E.), Computer Engineering — Gujarat Technological University (GTU), 8.6 CGPA

July 2023

CERTIFICATIONS

- Oracle Cloud Infrastructure 2024 Generative AI Professional — July 2024