Daksh Dhama

Phone: +91 9351217522 • Email: dakshdhama0502@gmail.com
LinkedIn: linkedin.com/in/daksh-dhama • GitHub: github.com/DaX-523

Summary

AI engineer (Python, TypeScript) who ships LLM systems to production at startups: RAG over embeddings and vector search, multi-agent workflows on LangGraph with checkpointed recovery, and real-time voice pipelines (STT, LLM, TTS). Delivered the video AI pipeline that converted a startup's first enterprise client, and cut per-call LLM cost **~77%** while tracking spend across 6+ model providers.

Skills

Languages: Python, TypeScript, JavaScript, SQL

Generative AI: LLM APIs (OpenAI, Gemini, OpenRouter), Prompt Engineering, Prompt/Context Caching, Tool Calling

RAG/Agents: RAG, Embeddings, Vector DBs (PGVector, Pinecone, Turbopuffer), LangGraph, LangChain, Multi-Agent Workflows

Voice AI: LiveKit, STT/ASR and TTS Pipelines, Multilingual Audio

Backend/Cloud: FastAPI, Node.js, Express.js, REST APIs, PostgreSQL, Redis, Celery, AWS (Lambda, SQS, S3, EC2), Docker

Tools: Langfuse, OpenTelemetry, PostHog, Pytest, Git, Cursor, Claude Code

Professional Experience

ButterCut AI **Remote**
Backend Lead & Full Stack Engineer *Aug 2025 – Present*

- Closed the platform's first enterprise deal by shipping a production-ready video pipeline to **SEEKHO**, converting it into a paying client and validating the product commercially.
- Built RAG retrieval pipelines (embeddings, vector search) that ground LLM generation in retrieved context for the platform's AI workflows, integrated with the FastAPI services and PostgreSQL data layer.
- Architected a five-stage async B-roll video pipeline (FastAPI, Docker) for parallel multimodal asset generation across overlay types (sticker, text, poster, video), backed by an AWS SQS worker with S3 offloading and containerized Lambdas at **99.9%** uptime.
- Reduced STT token usage by **89%** with a Two-Pass Architecture, reducing pipeline latency while preserving timestamp accuracy across multilingual audio.

Reachall AI **Remote**
Software Engineer (Part-time) *May 2025 – Aug 2025*

- Slashed post-call LLM analysis cost **~77%** (**$0.009** → **$0.002/call**) by removing duplicated prompt payloads, adding Gemini prefix/context caching, and fixing pricing bugs across cached-token and audio-token tiers.
- Built RAG retrieval into production voice AI agents, using embeddings and vector search to ground LLM responses in retrieved knowledge.
- Designed a sentence-level TTS response-caching layer (Redis/ElastiCache + S3, msgpack-serialized) that transparently wraps the TTS provider, eliminating redundant paid synthesis calls; rolled out platform-wide with per-agent toggles and cache-hit observability.

Hitori Tech **Remote**
Full Stack Engineering Intern *Dec 2024 – May 2025*

- Integrated the OpenAI Chat API for context-aware suggestions, raising form completion by **35%**.
- Developed backend systems and REST APIs for Clearance Resume, an AI-powered resume parser and builder, growing user sign-ups by **40%**.

Projects

AI UGC Ads Generation Pipeline
LangGraph, AWS Lambda, FastAPI *2025*

- Architected a 12-stage multi-agent video-ad pipeline on LangGraph state machines whose agent nodes call external generation tools (TTS, image-to-video, lip-sync, stitching), with per-node checkpointing for failure recovery and resumable, partial re-runs.
- Lowered generation cost **30–40%** via tiered asset-generation modes, disk-based asset reuse, and a real-time cost-telemetry system tracking spend across 6+ providers (OpenAI, Gemini/Vertex, Replicate, Sarvam, ElevenLabs, Pruna).
- Deployed as a serverless AWS Lambda/FastAPI service with a self-reinvoking worker pattern that checkpoints to S3, resuming multi-minute video jobs across Lambda's 15-minute execution limit.

Education

NIMS University **Jaipur, Rajasthan**
B.Tech in Computer Science Engineering *2021 – 2025*
GPA: 8.00/10.00 • Coursework: Data Structures and Algorithms, Cloud Computing, System Design