Om Anand

+91 77799 49969 | om.d3v.21@gmail.com | omanand0.in | linkedin.com/in/omanand10 | github.com/Om-anand-0
AI Engineer | B.Tech CSE 2027 | Bhopal, open to relocate

Education

Sagar Institute of Science & Technology (SISTec) — Bhopal, India
B.Tech, Computer Science & Engineering — *2023 – 2027 (expected)*

Experience

BrainChain Labs — Sep 2026 – Present
AI Engineer Intern — *Bhopal (Remote)*

- Building an AI operations platform for a US business client: a live operations dashboard and an owner-facing web app on Next.js, Supabase and the Claude API.

Freelance (client: YouTube creator) — Sep 2026 – Present
AI Engineer — *Remote*

- Building a 4-agent "chief of staff" with shared memory and a morning voice briefing (Vapi, ElevenLabs); RAG over sponsor files, human approval before anything publishes.

Projects

AI Digital Twin: voice RAG assistant | *Python, FastAPI, Qdrant, VoxCPM, Docker* — Mar 2026 – Present

- Built a voice RAG assistant: Groq Whisper speech-to-text, Llama 3.2 over Qdrant for grounded answers, and VoxCPM TTS fine-tuned with a custom LoRA voice adapter, driving a lip-synced Three.js avatar.
- Split inference into 3 Docker services so a TTS crash cannot take down the API; deployed on a hardened VPS (key-only SSH, Fail2Ban).

Analytics Q&A Agent: natural language to SQL for Slack | *Python, FastAPI, Instructor, SQLGlot* — Jul 2026

- Built a Slack agent that turns plain-English business questions into read-only SQL and posts tables, charts and cohort grids, using one structured LLM call (intent, SQL, chart, confidence).
- Replaced regex blocklists with SQLGlot AST guardrails (read-only, table whitelist, alias-aware columns) and a retry loop that feeds validator errors back to the planner.

Windtunnel: real-time GPU CFD in the browser | *TypeScript, WebGPU, Three.js, CI* — Sep 2026

- Implemented a lattice Boltzmann fluid solver in WGSL compute shaders, with GPU voxelization of any uploaded 3D model and FP16 storage to halve per-cell memory traffic.
- Built a validation suite that runs headless in CI against published references (cylinder Strouhal number 0.173 vs 0.166, sphere drag, Ahmed car body); interpolated walls more than halved sedan drag error (2.05 to 0.91).

Swasth AI: medical RAG assistant | *Python, FastAPI, Ollama, Qdrant, SSE* — Nov 2025 – Jan 2026

- Built a RAG backend that grounds answers in retrieved context and streams tokens over SSE, with consent-based tool calls and backend-enforced safety rules.
- Cut end-to-end latency from about 20 s to 6 s by bounding prompt size, gating retrieval and limiting memory injection.

Open Source

context-asr (PyPI, GitHub) | *Python, Whisper, spaCy* — Oct 2026

- Built a speech-to-text tool that detects spoken topics, pulls related names into Whisper's prompt and re-decodes, so rare names come out right; ships a CLI and an Earnings-21 benchmark harness.

BitChord (open-source YouTube Music client) | *merged PR #1134* — Oct 2026

- Replaced fixed sleep-timer presets with a custom 5 min to 12 h timer; merged in v1.8.2, shipped to users in v1.9.

Technical Skills

Languages: Python, TypeScript/JavaScript, C++, SQL, WGSL
AI / LLM: LLMs, RAG, Qdrant, embeddings, function calling, agents, prompt engineering, Claude/OpenAI/Gemini APIs, Ollama, Whisper, LoRA
Backend & Data: FastAPI, NestJS, Next.js, Supabase, REST, SSE, webhooks, OAuth, PostgreSQL, Redis, Pandas
DevOps & Tools: Git, Docker/Compose, Linux, Nginx, GitHub Actions (CI/CD), Vercel, VPS hardening, FFmpeg, WebGPU
CS Fundamentals: Data structures and algorithms, OOP, DBMS, Operating Systems