Umang Thakkar

Full-Stack AI Engineer | 14+ Production Systems Shipped | <500ms Voice Latency

umangthakkar005@gmail.com | +91 9426154668 | Delhi, India

LinkedIn: linkedin.com/in/umang-thakkar-90a4a5164 | GitHub: github.com/Umang00

Portfolio: umang-thakkar-full-stack-ai-engineer.vercel.app

Professional Summary

Full-Stack AI Engineer with 14+ shipped production systems. Achieved <500ms voice-to-voice latency with ElevenLabs + GPT-4o. Built MCP servers for Claude Desktop/ChatGPT. Expert in LLM fine-tuning, RAG pipelines, vector databases, and multi-agent orchestration. Hackathon winner (1st Runner-Up, Lecture Lens). Currently shipping production AI applications end-to-end.

Professional Experience

100x Engineers Cohort | AI Engineer

Remote | July 2025 – Present

- **Shipped 4 production AI systems in 6 months** including mobile apps, MCP servers, and multi-agent systems.
- **Built production MCP server (Git Roast)** enabling AI tool integration across Claude Desktop, ChatGPT, and Cursor with SSE streaming.
- **Developed multi-agent system (Breakup Recovery Squad)** using Agno framework with 4 specialized agents, and stateless privacy-first architecture, achieved viral adoption on Reddit.
- **Won hackathon for Lecture Lens**—RAG with hybrid ranking, pgvector, timestamp-aware chunking.

Hunch (Dating & Social App) | Associate Product Manager

Delhi, India | October 2023 – June 2025

- **Built AI voice agent** achieving <500ms voice-to-voice latency with interrupt handling and graceful fallbacks.
- **Fine-tuned GPT-4o on 450+ conversation pairs** with A/B testing framework, serving 100K+ DAU.
- **Built analytics dashboard** with Redshift SQL, Python pipelines, GPT-4o-mini for sentiment analysis on 1,000+ daily comments.
- **Designed MBTI matching engine** with Redis achieving sub-100ms response times for 100K+ DAU.
- **Built web onboarding funnel** with Next.js SSR, Stripe payments, Mixpanel analytics, and deep link handoffs.

Hunch (Anonymous Polling App) | Content Strategist

Delhi, India | November 2022 – September 2023

- **Built internal tooling** including Retool dashboards integrated with backend systems.
- **Designed data structures** for poll categorization and automated targeting logic.

PlotX (Crypto Gaming Platform) | Content Writer

Delhi, India | June 2022 – October 2022

- **Built content frameworks** driving 50% organic traffic improvement and 3,000+ monthly signups.

Key Technical Projects

AI Food Analyzer | Full-Stack Mobile App

React Native + Expo, FastAPI, Gemini 3 Pro (Multimodal), Neon Postgres

- Multimodal LLM analyzes ingredient photos directly to identify dietary concerns.
- 95%+ accuracy, shipped in 15 days as production-ready mobile app.

Lecture Lens | RAG System (Hackathon Winner)

Next.js, Supabase pgvector, OpenRouter, BM25

- Hybrid search (semantic + BM25), timestamp-aware chunking, clickable video citations.
- RLS multi-tenancy, built in 24 hours, won 1st Runner-Up.

Git Roast | MCP Server

TypeScript, Vercel Serverless, Gemini, GitHub API, SSE

- Production MCP for Claude/ChatGPT, SSE streaming, PDF generation for developer report cards.
- Achieved viral adoption on Reddit.

Fine-Tuned Chat Model | LLM Pipeline

Python, OpenAI Fine-tuning API

- 450+ curated examples, 3 training iterations, deployed serving 100K+ DAU with fallback logic.

Technical Skills

AI: LLM Fine-tuning, RAG (LangChain, LlamaIndex, Custom), Vector DBs (Pinecone, pgvector), Multi-Agent (LangGraph, Agno, OpenAI Agents SDK, Google ADK), Voice AI (ElevenLabs), MCP Servers, LoRA (ComfyUI)

Backend: Python, FastAPI, Node.js, Bun, Hono, PostgreSQL, Redis, Supabase, Serverless Functions

Frontend: React, Next.js, React Native, TypeScript, Tailwind

DevOps: Docker, Vercel, AWS, GCP, GitHub Actions

Education

B.Tech – Computer Science & Engineering

Parul Institute of Engineering and Technology | 2017 – 2021 | **GPA: 9.3/10.0**