

Chandan Kumar

Senior Software Engineer | AI / GenAI & Agentic Systems

📞 7233022562

✉ chandan.kumar.stack@gmail.com

in chandan kumar

🐙 github.com

📷 @ai_system_design

Summary

Senior software engineer (7+ yrs, JPMorgan Chase & BrowserStack) now building production AI — agentic developer tooling, low-latency voice agents, and LLM fine-tuning. Named BrowserStack's AI Champion (2025) for AI-native development; comfortable from GPU-level CUDA work to product delivery, with \$100K+ ARR and 5× TTS speedups delivered.

Education

IIT (BHU) Varanasi

B.Tech, Electrical Engineering, CGPA 9.03

July 2015 – May 2019

IIT Delhi (Distance Learning)

Quantum Computing & Quantum Machine Learning, Score 91/100

July 2022 – Jan 2023

Experience

BrowserStack

Mumbai

Senior Software Engineer | AI SPOC, Growth Eng

October 2024 – Present

- Created and drove org-wide adoption of a 6-phase agentic Spec-Driven Development pipeline (requirements → plan → tech-spec → test-cases → subtasks → implementation), standardizing idea-to-code across Growth Engineering — cutting development time ~50% and story bugs up to 70% on new initiatives.
- Built the “SDD Context Engine”, a Python FastMCP server exposing custom tools (feature-knowledge queries, dev-phase activation, mocking helpers) that injects codebase-grounded context into Claude Code / Copilot, sharply reducing model hallucination and drift.
- Owned the team's agentic AI harness — normalizing knowledge files, 11 skills, and 7 phase-specific agents (Requirements, Tech-Spec, Implementation, Bootstrap, Dev Setup) that auto-install on init; primary reviewer on dozens of harness/MCP PRs.
- Led an AI-driven Nightwatch → Playwright test migration across local/os1/os2/k8s environments, and shipped AI code review as a deploy gate with a cross-team PR approval/rejection skill.
- Served as AI SPOC for Growth Engineering — AI-SDLC enablement sessions, Claude Code training, and onboarding PMs and QAs onto the agentic workflow.
- Lead three expansion and retention initiatives (Free-Plan Access, Account Setup, Paywall), driving \$100K+ ARR; improved profile and pricing page load times 40–50%; mentor two engineers.

BrowserStack

Mumbai

Software Engineer

March 2022 – September 2024

- Built a self-serve promotional engine letting project managers launch ads or promotions on any BrowserStack page or flow.
- Designed complex upsell workflows for Pro plans across products and led a retention squad crafting personalized onboarding for new features.
- Built an end-to-end marketing-attribution system tracking UTM data and cross-page user activity, syncing it to Salesforce for the sales team.
- Integrated ZoomInfo enrichment across all site forms and built a custom payment flow for European countries affected by the Ukraine war.

JPMorgan Chase

Mumbai

Associate Software Engineer

March 2020 – March 2022

- Built a monitoring and alerting tool (Spring Boot + React) tracking 26 applications across different tech stacks — status checks, test reruns, and failed-test log inspection. Mentored an offshore team for ~1 year.

JPMorgan Chase

Mumbai

Software Engineer (SDE 1)

July 2019 – March 2020

- Built an end-to-end Spring Boot application for the SPG mortgage-analytics platform — the entry and exit point for all securitized-product analysis at JPMC.

AI & Independent Projects

Real-Time Voice Agent + Qwen3-TTS Megakernel Personal | github.com/chandan-123kumar/voice_stream

- Repurposed a persistent-kernel CUDA megakernel as the talker decoder for Qwen3-TTS (0.6B), delivering real-time streaming TTS at 47 ms median time-to-first-chunk and 0.145 RTF on an RTX 5090 (1033 tok/s decode) — 5× faster than the stock pipeline (0.738 → 0.148 RTF; first-audio latency 4.6 s → 47 ms).
- Built a full WebSocket voice agent (mic → local faster-whisper STT → gpt-4o-mini → megakernel TTS → speaker) on Pipecat with Silero VAD endpointing, barge-in, and 9 switchable voices; cut speech-end-to-first-reply latency 2–3.7× to 0.7–1.1 s by colocating STT on the GPU.
- Diagnosed and fixed a grid-barrier race deadlock in the decode kernel; verified numerical parity against the HF reference (per-layer bf16-noise match, 0.98 hidden-state cosine).

Indic LLM Fine-Tuning Personal

- Fine-tuned open-source LLMs (Qwen) for Indic languages using QLoRA; evaluated tokenizers and compared SFT vs. continued pretraining.

@ai_system_design Technical Content

- Create AI system-design education for software engineers on Instagram and YouTube — multi-agent patterns, LangChain / LangGraph / CrewAI, distributed systems, and database internals.

Achievements

- AI Champion (Dec 2025) — company-wide recognition from the CTO and Engineering Director for establishing AI-native development practices across Growth Engineering; 10+ peer and leadership awards for innovation, deep-dive debugging, and cross-team collaboration.
- HackerRank: Silver — Week of Code 37; Bronze — World CodeSprint 13.
- 2nd place, La-Trajectoires, Technex'16 (IIT BHU technical festival).

Technical Skills & Certifications

- AI / GenAI: LLM fine-tuning (QLoRA, SFT), RAG, multi-agent systems, agentic SDLC, AI code review, prompt & context engineering, agent evaluation
- Agent Frameworks: LangChain, LangGraph, CrewAI, MCP / FastMCP (Model Context Protocol)
- Voice AI: Pipecat, Twilio, ElevenLabs, Cartesia, Whisper / faster-whisper (STT), Silero VAD, low-latency TTS (Qwen3-TTS)
- Languages: Python, Java (Spring Boot), JavaScript / React, C, Bash, SQL
- Systems & Infra: CUDA / GPU programming, FastAPI, Docker, PostgreSQL, Redis, Kafka, MongoDB, AWS
- Certifications: Advanced Java, Problem Solving (Advanced), Data Structures & Performance, OOPs in Java, Quantum Computing & ML — IIT Delhi, HTML, CSS, JavaScript