

VANSSH AGARRWAL

+91-9609903555 · vansshagarrwal@gmail.com · github.com/code-1705 · linkedin.com/in/vanssh-agarrwal · Bengaluru, India

SUMMARY

Software engineer who builds AI agents that run in production: tool use, RAG, multi-agent orchestration with LangGraph, evals and per-call tracing, on Python/FastAPI and AWS. Hands-on with payments (agentic collections, Razorpay reconciliation, payment-code safety rules) and real-time voice and video inference.

TECHNICAL SKILLS

AI Agents & LLMs: LLM APIs (Anthropic, Gemini), LangGraph (tool use, RAG, multi-agent), MCP, prompt design, evals, tracing, vector databases (Pinecone, pgvector), Claude Code

Multimodal & Voice: Whisper (STT), Coqui TTS, YOLO, FaceNet, WebRTC, real-time audio/video streaming

Languages: Python, Go, TypeScript/JavaScript, Java, C++, SQL

Backend & Distributed Systems: FastAPI, REST, WebSockets, SSE, Microservices, Redis (caching, pub/sub, sliding windows), PostgreSQL, MongoDB, MySQL, DynamoDB, webhooks, background jobs

Cloud & Tools: AWS (EC2, S3, DynamoDB, IAM), Docker, Docker Compose, Kubernetes, CI/CD, Git, PyTest (unit & integration testing)

WORK EXPERIENCE

OpenIntervue | Software Development Engineer Intern | AI hiring startup Oct 2025 – Present

| Python · FastAPI · Anthropic Streaming API · LLM APIs · Claude Code · Notion · WebRTC · WebSockets · YOLO · FaceNet · MongoDB · Redis · Docker · AWS

- **Pihu Meet, AI agents in a live interview room:** build backend and frontend for OpenIntervue's live AI interview product. Its AI notetaker transcribes and suggests follow-up questions to recruiters; its AI panelist, PIHU, co-interviews and takes over if the human panelist drops. Every LLM and tool call is logged against the interview, so a bad answer in a live room can be traced step by step afterwards.
- **Real-time multimodal inference:** built a proctoring service in Python/FastAPI that runs face and object detection (FaceNet, YOLO) on live interview video and tracks each session through a state machine, keeping round-trip latency under 180ms.
- **LLM pipeline at scale:** rewrote the LLM resume-parsing pipeline to stream responses with chunked delivery and connection pooling, cutting p99 latency by ~70% and handling 1,000+ concurrent parses. Prompt and model changes are run against an eval set of labelled resumes before they ship.
- **Autonomous issue-triage agent:** set up an hourly agent that picks up open candidate issues from the team's Notion board, pulls that candidate's logs, traces the root cause across our codebases and writes the finding back to Notion.
- **Reliability:** fixed production failure alerts that never fired across 8 server workers by moving the counter to a shared Redis sliding window, so the team is paged as soon as an endpoint starts failing. Wrote a PyTest integration suite covering 50+ API endpoints and 25+ race-condition scenarios.

PROJECTS

Resolve.ai — Multi-Agent Payment Collections Platform · github.com/code-1705/ResolveAI

| Python · FastAPI · LangGraph · PostgreSQL · Pinecone · Razorpay · SSE

- Built a FastAPI backend where LangGraph agents classify customer intent, apply payment rules and settle invoices, with agent state in PostgreSQL. Agents use tools, follow policy and change backend state.
- Reconciled payments automatically through Razorpay webhooks, grounded agent replies in policy documents with retrieval over a Pinecone vector index, and pushed live updates to the client over SSE.

SpeechBridge — Real-Time Voice Gateway for AI Agents

| Go · Python · WebSockets · Whisper · Coqui TTS · Redis Pub/Sub · Docker

- Built a full-duplex WebSocket gateway in Go that streams audio between users and AI agents, with self-hosted Whisper speech-to-text and Coqui text-to-speech, at under 250ms round-trip latency.
- Moved session state into Redis with pub/sub so any gateway instance can serve any session, letting the service scale horizontally across Docker containers.

FinGuard — Payment-Code Reviewer · AIM Code Kitchen S01 (Google Cloud) hackathon · github.com/code-1705/Codacy

| Python · AST Parsing · Static Analysis

- Built and demoed deterministic guardrails for payment code: AST rules that flag float currency math, network calls inside DB locks, and missing idempotency keys.

CivicFix — Civic Complaint Tracking Platform on AWS · civicfix.duckdns.org · github.com/code-1705/CivicFix

| Python · FastAPI · Next.js · AWS (EC2, DynamoDB, S3, IAM) · Docker Compose

- Built, deployed and run a live platform on AWS EC2 via Docker Compose with complaints in DynamoDB, photos in S3 and least-privilege IAM roles; duplicate reports merge by location and a background job escalates tickets past their 24–72h deadline.

EDUCATION & CERTIFICATIONS

Bangalore Institute of Technology

2023 – 2027

Bachelor of Engineering in Information Technology **CGPA: 8.33/10** · No active backlogs

Certifications: Machine Learning with Python (IBM); GitHub Foundations (GitHub) · **Problem solving:** 200+ DSA problems on LeetCode, CodeChef and GeeksforGeeks