

Akshit Ahuja

AI Engineer · Agents in production: evals, reliability, voice · Python, TypeScript

ahujaakshit20@gmail.com · +91 70877 55262 · [Linkedin](#) · Remote, India (IST; 3–4h daily overlap with US hours) · Contractor or EOR

AI engineer and technical co-founder. I take agents from prototype to production and keep them working: human-in-the-loop gates, confidence routing, rubric scoring, idempotent pipelines, retries and cost control. 4 years of Python at scale (**34M+ requests/day** at Spinny), then founding engineer at PeerSeek, where I built the whole backend and ran sales calls and user interviews myself. Since then I have shipped LLM extraction that took a **20-person** manual process to **3**, multi-agent pipelines on LangGraph and the OpenAI Agents SDK, Hindi-first voice agents on Pipecat, and MCP servers. I work embedded with the customer: scope the messy problem, ship the agent, own the outcome.

EXPERIENCE

PeerSeek.io

Remote

Co-Founder & Founding Engineer

May 2025 – Present

- **Multi-agent content pipeline (Folio, tryfolio.site).** Plan → research → outline → draft → score → revise → **human approval** → publish, each stage a Celery task. An LLM rubric-scoring stage sends weak drafts back to revision with notes; every claim must trace to recorded evidence; site facts stay *pending review* until a human confirms. Model routing (strong model writes, cheap models do grunt work) on the OpenAI Agents SDK over Azure Foundry; browser-extension bridge for agent research; multi-tenant; 100+ test modules.
- **Platform, built solo.** The entire backend of a creator-monetization platform: **18 Django apps** (bookings, webinars, DRM digital products, payments and payouts, analytics, search, realtime chat), on Kubernetes (Azure AKS) as separate web, WebSocket, worker and scheduler deployments.
- **One payment engine for every product.** Polymorphic offering pointer shared by all money tables: Razorpay orders and webhooks, per-component pricing, stacking coupons, commission, and a held→settled ledger driving payouts. New product types inherit payments without touching the money path.
- **Event-driven automation.** Instagram auto-DM on the Meta Graph API with idempotent webhook handlers, follow-gating and token refresh; email and WhatsApp notifications de-duplicated by message ID.
- **Customer-facing.** Ran early sales conversations, user interviews and onboarding; turned what users asked for into the roadmap.

Independent AI Engineer

Remote

Embedded with clients to scope, build and hand over production AI systems

2025 – Present

- **LLM document extraction (billion-dollar company).** Connector for a vendor portal with no API, document parsing and LLM field extraction, with **low-confidence reports routed to a human review queue**. Replaced a **\$10,000/month** manual process and took the team from **20 people to 3**.
- **Voice agents.** Hindi-first AI calling platform split into three planes (Django control plane, FastAPI + Pipecat voice plane, analysis API) with telephony/LLM/STT/TTS adapters whose boundaries import-linter enforces in CI (ruff, mypy strict, pytest gate). Built an outbound calling agent on Twilio Media Streams with voicemail-vs-human detection and IVR navigation via DTMF, and designed the turn pipeline (VAD, streaming STT, smart-turn, sentence streaming to TTS) against a ~600ms end-to-end latency budget.
- **Call intelligence (Happy Nature).** Transcribes sales calls, scores agents against a rubric, flags red flags, and answers questions over call data with text-to-SQL and pgvector RAG; benchmarked four STT providers on real Hindi/Hinglish calls before choosing one.
- **Autonomous QA agent.** Multi-agent LangGraph system (planner, executor, verification, visual regression, test generator) that explores an app on a real device, finds bugs and writes regression tests; outputs validated against real flows for hallucinations.
- **Infra.** Self-hosted LLMs on Azure GPU VMs behind a wake-on-request gateway (near-zero idle cost); MCP servers for internal tools.

Spinny

Gurgaon, IN

Software Engineer Intern → Software Developer → Software Development Engineer 2

Aug 2022 – Apr 2025

- **Dealer Finance.** Partnered with the central finance team and built the entire integration for the dealer lending system, which disburses **₹30 crore+ (~\$3.4M)** in loans every month.
- **Connected Cars Platform.** Backend handling **34M+ daily requests** on MongoDB, Redis, AWS Lambda and Elastic Beanstalk.
- **Auction Scheduler.** Event-driven microservice on Celery and Kafka automating **35k+ vehicle auctions**.
- **Central Ticketing.** Built the auction vertical's ticketing on the company-wide ticketing system after mapping dealer queries with business and product; resolution time down **30%**.
- **Reliability at scale.** Led the migration from self-managed RabbitMQ to AWS SQS; idempotent payment-webhook APIs with retries and exponential backoff cut external API failures by **90%**; New Relic, Sentry and gateway alerting for **98%+ uptime**.
- **Performance.** Introduced gevent on services fanning out to third-party vendors, reducing resource usage by **30%**; MongoDB indexing took an API from **500ms to 150ms**.
- **Cross-team delivery.** Engineer of choice for cross-pod integrations and new business initiatives.

Zscaler

Chandigarh, IN

Cloud Operations Intern · Python and Bash automation for access control and cloud monitoring

Feb 2022 – Jul 2022

SKILLS

Agents	LangGraph, OpenAI Agents SDK, Claude Agent SDK, MCP servers, tool/function calling, structured outputs, multi-agent orchestration, model routing
Eval & reliability	Rubric (LLM) scoring gates, confidence routing to human review, claim/evidence provenance checks, provider benchmarking on real data, idempotency, retries and backoff, observability (Sentry, New Relic, Prometheus/OpenTelemetry)
Voice & retrieval	Pipecat, Twilio, Plivo, Deepgram, Sarvam, ElevenLabs, Silero VAD; pgvector RAG, text-to-SQL
Models	Claude, GPT, Gemini via Anthropic, OpenAI and Azure OpenAI APIs; self-hosted models on GPU VMs; Claude Code for daily AI-augmented development
Languages & backend	Python, TypeScript, SQL, Bash; FastAPI, Django/DRF, Channels, Celery, Kafka, WebSockets; Next.js
Infra & data	AWS (Lambda, Elastic Beanstalk), Azure (AKS, Blob, GPU VMs), Docker, Kubernetes, CI/CD; PostgreSQL, Redis, MongoDB

EDUCATION

Guru Nanak Dev Engineering College — B.Tech in Computer Science, CGPA 8.02

Ludhiana, IN · Aug 2018 – Jun 2022

Bunk Master: built an attendance planner used by **500+ monthly active students** before COVID · backend of the college fest website · Convenor, SCIE society (ran a coding contest).