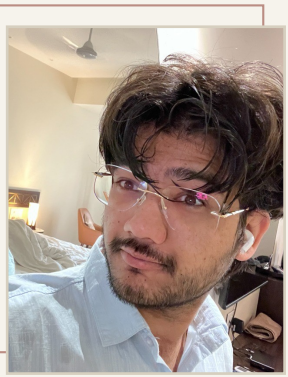

Rahul Soni.

AI and automation that save businesses time and money.

Self-taught builder and architect. I put AI and automation to work inside businesses: new products, sharper versions of what they already run, and automation that removes expensive manual work, so they save thousands of hours and serious money. I'm the one you bring in when it gets complicated. I own the whole thing, problem to shipped product, then stay on to lead the team that scales it.

irahulsoni.com · rahul@only4values.com · linkedin.com/in/irahulsoni
Works internationally · the go-to when it gets complicated

01 What I've shipped — for enterprise

Described by function (clients and employer stay private). The shape of the impact is consistent: thousands of human hours automated, six-figure ($) savings in the first months, and value that compounds into the millions as the system runs. Not the whole list — more is in build, and more isn't public yet.

AI contract-intelligence platform — shipped
Extracts the clauses that matter, flags risk, and runs review work once done by hand, across tens of thousands of documents in production. Saving six figures ($) in costs from its first months, and still compounding.

Intelligent document-extraction platform — flagship
Turns the messiest real-world documents into clean, structured data; multi-tenant, live in production. 97–98% accuracy across client configurations, up from a ~60% legacy baseline. Onboarding cut from months to days. Architected for millions of documents. Treated by leadership as an emerging revenue line.

Multi-language media & social platform — founder build
Social-media management across platforms and languages, with a real media pipeline at its core: transcription, translation, and voice dubbing. My own venture, built end to end. Brief on purpose.

Proposal-automation engine — shipped
Retrieval over a library of historical proposals plus semantic search, so responses to complex requests get drafted in context, grounded in prior work instead of a blank page.

High-throughput ingestion pipeline — shipped
A streaming, staged-concurrency redesign of a document-processing path. About 30–40x the throughput of what it replaced, on the same hardware.

02 Built independently

Built because I wanted it to exist. The kind of infrastructure companies fund whole teams to attempt; I build it alone and use it every day.

Yantra — AI development environment
Built from scratch to keep agents, projects, and context coherent across dozens of parallel workstreams. A full, signed, notarized app: 600+ commits, 900+ tests, my daily driver for months. Not a plugin, not a fork.

Stenograph — private knowledge engine
Deep research grounded in your own context and projects, with an MCP layer that hands it straight to the AI agents I run. On track to process well over a million hours of media across every domain by late 2026.

Real-time voice-AI agent
A telephony agent that holds a live conversation: streaming speech-to-text into an LLM and back to voice, with voicemail detection and barge-in, under 500ms end to end.

Large-scale video-understanding R&D
Self-funded research into understanding media at scale: tens of thousands of items ingested with near-zero failures, vision-language models run across a 9x H100 GPU fleet, 14 models benchmarked across 250+ experiments, 6–7x latency cut by datacenter engineering.

---

03 Applied & agentic AI

My deepest area, and the engine behind every product I ship.

Early and continuous
On this since the first wave (ChatGPT, GitHub Copilot) through Anthropic's Claude Code and its Max tier. Usually first to new tooling, and the one who helps the team adopt it.

Daily, at scale
Parallel AI agents are my normal way of working. I instrument my own usage, thousands of sessions of it, to keep getting sharper.

Across the whole stack
RAG pipelines, MCP servers that hand agents real tools, agentic workflows, vision-language models, local and open models (Ollama), and the wider ecosystem.

---

04 Skills & capability

The technical surface, drawn from real production work, not a checklist.

AI & agentic (deepest area)
LLM orchestration (Gemini, OpenAI, Claude), RAG, MCP (building & exposing tools to agents), agentic workflows, prompt engineering, vision-language models, model evals, local/open models.

Backend & APIs
FastAPI (async Python), NestJS / Node.js, SQLAlchemy, REST / WebSockets / Pub-Sub, Temporal (durable workflows), Redis.

Frontend & native
React / Next.js, TypeScript, Tailwind / shadcn; React Native / Expo (mobile); Swift / SwiftUI signed & notarized macOS apps.

Cloud & infrastructure
AWS / GCP / Azure (cloud-independent), Docker, Kubernetes (AKS) / Helm, Terraform / IaC, self-hosted CI, GPU fleets (H100 / A100 / L40S).

Data & ops
Terabyte-scale pipelines, ETL / BI, vector DBs (pgvector / Qdrant / Chroma), PostgreSQL / MongoDB / Redshift, media (FFmpeg, STT, TTS); Grafana observability, cost optimization.

Roles & how I've worked
Builder, architect, technical lead, advisor; often more than one at once. Brought in and trusted as a technical lead and senior architect. Solo end to end, embedded with teams, leading delivery, and as a founder, for clients from small businesses to enterprise.

05 Trajectory

2019
Taught myself to code from a non-technical background. No degree, no bootcamp.

Soon after
Scaled to a $100–200/hour rate and $15K+ months, delivering for international clients who kept handing me harder work, then the delivery itself.

Founder
Built and ran my own startup on my own capital. I learned what a P&L feels like from the inside, and in my own ventures I build the team and lead it.

Compounding since 2019
25,000+ hours of continuous work. Not classroom hours: shipped projects, paid work, and relentless leveling across new domains and tools.

Now
Shipping enterprise-grade AI that's chased by multi-million-dollar deals.

06 Backed to build

Program backing: credits, compute, and tooling, not raised investment. Company after company looked at the work and chose to back it.

Over $350K in backing, across a dozen+ programs
A combined pool worth more than $350,000 in credits, compute, tooling, and SaaS from AssemblyAI, Microsoft for Startups (Founders Hub), Google for Startups Cloud, AWS Activate, NVIDIA Inception, OVHcloud, Modal, Daytona, Atlassian for Startups, ElevenLabs, Google Workspace, and more.

My own capital on top
Well into six figures ($) of my own money in hardware, compute, and the R&D behind the work. Approaching half a million dollars of resources behind the work, all told, over the past few years.