

# Parik Ahlawat

## Senior AI Consultant & Engineer | Autonomous Agents, Multi-Model Orchestration & Production AI Systems

pa.parikahlawat@gmail.com | Remote | linkedin.com/in/parik-ahlawat | github.com/par1kahl

### PROFESSIONAL SUMMARY

Senior AI Engineer and Consultant with 6+ years building production AI systems. **618+ automations shipped across 35 projects in 12 industries** — AI/ML infra, advertising, sales/CRM, media/creative, fintech/trading, e-commerce, developer tools, marketing SaaS, logistics, interior design, and security. I design and ship autonomous agent architectures, multi-model routing platforms, and AI operating systems that replace fragmented SaaS tools.

Creator of **Kronus** (open-source agent orchestration framework — 21 agents, 27 skills, swarm mode, Telegram control plane daemon, multi-tenant hub with JWT + Stripe billing) and Founder of **Phavella** (AI business automation platform spanning marketing, sales, finance, and e-commerce ops). Systems have improved LLM accuracy 15-25% on industry benchmarks, reduced developer effort ~60%, and maintained 99.9% uptime in production. I specialize in AI that works at 3am without paging anyone.

### CORE COMPETENCIES

Consulting & Strategy	Engineering & Architecture
AI system architecture reviews	Multi-agent orchestration & swarm mode
Production readiness audits	RAG pipeline design (Qdrant, pgvector, ChromaDB)
Cost optimization & model routing	Full-stack AI apps (FastAPI, Next.js 15, PostgreSQL)
Requirements to production planning	LLM evaluation (BFCL, TAU-bench, LiveBench)
AI integration into existing stacks	MCP server development (34-tool servers shipped)

### TECHNICAL SKILLS

**AI / ML & Agent Infra:** LangChain, LangGraph (agentic RAG + Text-to-SQL), OpenAI, Anthropic Claude (Haiku/Sonnet/Opus), Google Gemini, Mistral, Llama | Multi-agent orchestration, swarm mode, MCP (Model Context Protocol), tool-calling, Celery, Bull, node-cron | Neurosymbolic reasoning, Small Reasoning Models (SRMs) | Evaluation frameworks (Gorilla BFCL, TAU-bench, LiveBench)

**Backend:** Python, TypeScript, FastAPI, Django Ninja, Node.js, Hono, Express, REST, GraphQL

**Frontend:** React, Next.js 15, Tailwind CSS, Three.js, Recharts, Tremor, shadcn/ui, Vite, XYFlow, Framer Motion, GSAP, Zustand, TanStack Query

**Data & Storage:** PostgreSQL, MongoDB, Redis (ioredis), SQLite, Supabase, SQLAlchemy, Alembic, Prisma, Drizzle ORM | Qdrant, Pinecone, ChromaDB, Neo4j, pgvector

**Automation & Integrations:** n8n (4,274 workflows catalogued), Meta Ads, Shopify, Slack SDK, Telegram Bot API, Google Sheets/Drive API, Zoom, Zerodha, Plaid, Stripe, fal.ai, ElevenLabs, Apify, Read.ai

**DevOps & Monitoring:** Docker, Kubernetes, GCP, AWS, Render, GitHub Actions, Dependabot, PM2, nginx, Sentry, Playwright, Zod



## PROFESSIONAL EXPERIENCE

---

### Founder & AI Architect | Phavella

2026 – Present | Remote

*Building a suite of AI operating systems that replace fragmented SaaS tools with unified, self-running platforms for businesses.*

- **mediaOS** — End-to-end video ad production pipeline: 13-state FSM, Gemini forensic analysis, 11 video models via fal.ai, ElevenLabs voiceover, Meta Ads auto-upload with winner/loser classification. Deployed and live.
- **salesOS** — AI sales intelligence for founder-led teams: zero-admin CRM, meeting transcription, deal coaching via Claude, RAG sales chat with MEDDIC/SPIN/BANT frameworks.
- **finOS** — AI finance OS: runway engine, receipt OCR (99.8%), Prophet/ARIMA forecasting, investor update auto-gen, tax compliance RAG.
- **ecommerceOS** — 12-state product pipeline: AI descriptions, SEO, dynamic pricing, demand forecasting, multi-channel sync (Shopify/Amazon/WooCommerce/Etsy).
- Multi-model orchestration engine routing tasks to optimal models (Claude Haiku/Sonnet/Opus, Gemini, OpenAI).

Stack: FastAPI, Next.js, React, PostgreSQL, Redis, Celery, Qdrant, pgvector, Claude, Gemini, fal.ai, ElevenLabs, Plaid, Stripe, n8n

### Product Manager + Senior AI Engineer | CraftifAI

Oct 2025 – Dec 2025 | Remote

*Led AI agent product for embedded and edge platforms where cloud-based models are not an option.*

- Designed agent workflows operating within real hardware constraints (memory, latency, connectivity).
- Built model orchestration layer selecting between on-device and cloud inference based on resource availability.
- Managed product roadmap and engineering execution simultaneously.

### Founding Senior AI Engineer | CoreThink AI

Mar 2025 – Aug 2025 | Remote

*First engineering hire. Built LLM evaluation and reasoning infrastructure from scratch.*

- Designed unified evaluation framework integrating Gorilla BFCL, TAU-bench, and LiveBench with common base\_url abstraction for parallel multi-model testing.
- Built neurosymbolic Chain-of-Thought pipeline: NL processing, symbolic representation, inference engine, and 4-layer verifier (structural, logical, cross-check, failure detection).
- Developed Small Reasoning Models (SRMs) — lightweight verifiers that run inline in inference pipelines.
- **Impact: LLM accuracy +15-25% on SWE-Bench, BFCL, TAU, LiveCodeBench. Agent latency -30% via batched vector search and caching.**

### AI Engineer & Researcher | AccelChain

Apr 2023 – Aug 2025 | Remote

*Built AI-powered developer tools for automated code analysis, generation, and security auditing.*

- Built multi-stage AI code generation pipeline: task normalization, RAG-style context injection, multi-pass LLM generation, validation layers, iterative refinement loops.
- Designed 4-agent coding workflow: planner, coder, reviewer, fixer with controlled iteration loops and cross-agent state.
- Built automated audit agents with self-critique loops; Solidity AST parser with 5-scan security engine.
- **Impact: Developer effort -60%. Two foundation grants secured for AI tooling research.**

### Freelance Engineer | Sidechain Labs

2019 – Present | Remote

*Independent consulting practice. Backend systems, APIs, AI tooling.*

- Delivered 20+ client solutions across backend architecture, API design, and open-source AI tooling.
- Collaborated with engineering teams at Google and other major tech companies.



## KEY PROJECTS

---

### Phavella | AI Business Automation Platform

End-to-end business automation suite replacing fragmented SaaS stacks with unified, AI-native workflows across marketing, sales, finance, and e-commerce operations. Multi-model orchestration engine, fault-tolerant FSM pipelines, RAG knowledge bases, and Slack-driven human review loops automate the full ops stack — from ad creative production and Meta Ads upload, to CRM-less sales intelligence, receipt OCR + runway forecasting, and multi-channel product pipelines. Production deployments live.

Tech: FastAPI, Next.js, PostgreSQL, Redis, Celery, Qdrant, pgvector, Claude, Gemini, fal.ai, ElevenLabs, Plaid, Stripe, n8n, PM2, nginx

### Kronus | Open-Source Agent Orchestration Framework

Two-part product: **Kronus Core** — hybrid agent + skill architecture with 21 specialized agents, 27 skills, swarm mode for parallel multi-agent teams, model routing (Claude Opus/Sonnet/Haiku), MCP server integrations, Connect SDK, 7-page React dashboard (sessions, knowledge graph, logs, security, todos, features), and a multi-tenant Hub server with JWT auth + Stripe billing (108 automations). **Kronus Daemon** — Telegram control plane routing messages to per-project Claude Code sessions, with persistent session manager, domain classifier, context overflow auto-reset, persona management, collaborator mode, journal + memory systems, and 14-page dashboard (42 automations).

Tech: TypeScript, Bun, Hono, React, Vite, Tailwind CSS, D3.js, Recharts, PostgreSQL, Redis (ioredis), SQLite, Drizzle ORM, Jose (JWT), Argon2, Stripe, Grammy, MCP SDK, Commander.js, Playwright, Docker

### Creative OS | AI Ad Creative Platform (Client)

522 commits. Five integrated sub-systems: P1 Ad Intelligence (competitor scraping via Apify), P2A template-based generation, P2B native generation, P2C hook generation, P3 video pipeline with avatars, B-roll RAG, scene workbench. Central COS control panel with PM2 management.

Tech: Next.js, FastAPI, PostgreSQL, Qdrant, Apify, PM2, Meta Ads API

### CommsOS | RAG Meeting Intelligence Platform (Client)

Ingests meeting transcripts from Read.ai, Zoom, and Google Calendar; chunks and embeds in Qdrant; AI search/ask/chat with intent detection. Action item tracking with Slack delivery, people profiles, streaming agent chat endpoint. Docker + nginx deployment.

Tech: FastAPI, Next.js, Qdrant, PostgreSQL, Slack SDK, Zoom API, Read.ai, Docker

### statAISA | Visual Trading Strategy Builder (Client)

93 commits. Visual node-based strategy editor (XYFlow) with backtesting engine, live market data via WebSocket, Zerodha broker integration, payment system, 2FA auth, Sentry monitoring, CI/CD with Dependabot.

Tech: FastAPI, Next.js, PostgreSQL, XYFlow, Zerodha API, Sentry

### metaAds-mcp | MCP Server for Meta Ads Intelligence (Client)

Production MCP server exposing 34 tools: 26 marketing (account overview, campaign insights, demographics, budget, alerts) + 8 competitor intelligence (active ads, history, creative texts, spend summary, reach). PostgreSQL-backed, Apify ingestion, auth middleware, rate limiting.

Tech: TypeScript, MCP SDK, PostgreSQL, Apify, Zod, Vitest

### Besque | AI Video Generation Platform (Client)

Full-stack AI video generator with context bank, reference library, style profiles, Gemini vision analysis, fal.ai Seedance generation, prompt history, Celery job queue, Google Drive cloud sync.

Tech: FastAPI, React, PostgreSQL, Redis, Celery, Gemini, fal.ai, Google Drive API

### SalesLeadNurture AI Agent (Client)

Agentic RAG built with LangGraph: multi-node graph (router → RAG → Text-to-SQL → response), campaign management with AI-generated messages, conversation intelligence with intent detection, analytics dashboard, evaluation suite. Docker-deployed on Render.

Tech: Django Ninja, React, LangGraph, pgvector, Docker



## KEY ACHIEVEMENTS

---

- **618+ automations** shipped across 35 projects in 12 industries
- **392 tools** integrated across the production stack
- **Kronus** — 21 agents + 27 skills + swarm mode, 150 automations across Core and Daemon
- **Creative OS** — 522-commit client ad platform, 5 sub-systems
- **34 MCP tools** in a single Meta Ads server
- **15-25% LLM accuracy improvement** across 4 benchmarks
- **60% reduction** in developer effort via AI workflows
- **30% latency reduction** via batched vector search
- **4,274 n8n workflows** catalogued and productionized
- **99.8% receipt OCR accuracy** in finOS
- **Two foundation grants** for AI tooling research
- **100% Job Success** on Upwork

## EDUCATION

---

**B.S. Computer Science** | Purdue University, West Lafayette, IN, USA | 2018