Deep Kasajoo

20+	$225K	45	10×	20+
Paying Restaurants	Proj. MRR by 2027	Days CafeOS Built	Developer Velocity	Yrs Enterprise

AI Engineer · LLM Orchestration · Systems Architect · UI/UX & Creative Director · Full-Stack Developer · Founder

dip@kasaju.com · kasaju.com · cafeos.co · Dallas–Fort Worth, TX

PROFESSIONAL SUMMARY

Veteran systems architect with 20+ years of enterprise experience and an AI-native engineer since 2021. Founder of Gravity Technologies and CafeOS — a live, revenue-generating, zero-commission restaurant OS with 20+ paying clients, built end-to-end in 45 days using agentic AI orchestration. Applies disciplined model selection — continuously routing tasks across Claude, GPT-4o, Gemini, DeepSeek, and specialized tools based on task type, cost, latency, and each model's evolving strengths, updating that judgment throughout the day as the AI landscape shifts. Delivers platforms that are enterprise-grade in architecture, consumer-quality in experience, and production-ready from day one.

AI ENGINEERING & PRODUCT DEVELOPMENT · 2021–PRESENT

Founder, AI Architect & Lead Engineer

Gravity Technologies (Operating under DFW Branding LLC) · Dallas–Fort Worth, TX · 2021–Present

- Built CafeOS — a 12-module, zero-commission enterprise restaurant OS — end-to-end in 45 days using agentic AI orchestration (GCP, Claude Code, ChatGPT Codex, Claude Sonnet/Opus Thinking, GPT-4o, Gemini 3.1 Pro, DeepSeek R1, Kimi, Manus, Nano Banana). Now serving 20+ paying restaurants across TX, CA, and IL at $100–$350/month, targeting 1,000 restaurants and $225K MRR by end of 2027.
- Engineered the GenieUs AI agent system: operators update their live website by sending an SMS, email, or WhatsApp message in plain language — no backend access needed. Client-side GenieUs assists customers with cart management, food selection, and order completion. Backend GenieUs delivers real-time analytics via natural conversation (revenue today, top menu items, monthly trends).
- Launched an industry-first TV Kiosk Ordering System within CafeOS — a feature not available in any competing restaurant SaaS platform at launch.
- Designed LLM orchestration pipelines routing requests across Claude Opus/Sonnet (Thinking), GPT-OSS 120B, Gemini 3.1 Pro, and DeepSeek R1 based on task complexity, cost, and latency. Semantic caching reduces API costs 40–60% in production.
- Built multi-tenant RAG pipelines (pgvector) with RAGAS evaluation — monitoring faithfulness, answer relevancy, and context precision. Tracked p95 latency and TTFT for all streaming AI endpoints.
- Engineered multi-tenant row-level security in Supabase/PostgreSQL with strict tenant isolation, ADA compliance, and full security audits — deployed and verified in production.
- Deployed CI/CD pipelines on GCP (Cloud Run, Cloud Build, Artifact Registry) with blue-green deployments, rollback automation, and observability via LangSmith and Helicone.
- Designed all UI/UX from scratch in Figma — component systems, design tokens, responsive layouts — resulting in platforms that are consumer-quality in experience and enterprise-grade in architecture.
- Provided AI consulting and implementation services for government economic development bodies (Lavon EDC, Sanger TX, Crawford City TX), independent businesses, and freelancers — advising on AI adoption strategy, workflow automation, and platform selection.

ENTERPRISE ARCHITECTURE & LEADERSHIP · 2000–2021

Creative Director & Technology Director

Dreams & Ideas · Nepal · 2000–2017

- Led end-to-end technology architecture and creative direction for 50+ clients spanning SMBs, multinational corporations, and government agencies — delivering CMS platforms, web applications, desktop software, and digital infrastructure.
- Designed and built the CMS and web platform for the National Information Agency of Nepal, and architected the official digital platform for the President of Nepal, H.E. Bidya Devi Bhandari.
- Designed UI/UX for banking ATM systems — applying UX research, accessibility principles, and visual design to mission-critical financial hardware interfaces.
- Architected and operated Nepal's #1 ranked website for 5+ consecutive years — a high-traffic entertainment and information portal built and scaled entirely in-house.
- Led and managed a cross-functional team of 20–30 engineers, designers, marketers, and content creators over 5 years (2007–2013), delivering complex digital platforms on time and at scale.

Founder & Creative Technology Director

DFW Branding LLC · Dallas–Fort Worth, TX · 2017–Present

- Founded and operate a DFW-based creative and technology consultancy delivering brand identity, photography, video production, web design, digital marketing, and technology architecture to businesses from independent professionals to government bodies.
- Parent entity of Gravity Technologies — the AI engineering and systems development division handling all consulting, product development, and enterprise technology work from 2021 onward.
- Clients span the DFW metro and international markets — branding, creative production, and digital strategy running in parallel with AI product development since 2021.

TECHNICAL SKILLS

LLM Orchestration
· Multi-agent pipelines
· Context routing
· LangChain / LlamaIndex
· AutoGen · CrewAI · MCP

Agentic Workflows
· Autonomous execution
· Tool-use & function calls
· LangGraph · Semantic Kernel
· Prompt scaffolding

RAG & Vector DB
· Embeddings & chunking
· Pinecone · pgvector
· Hybrid dense/sparse
· RAGAS · p95 evaluation

Prompt Engineering
· System-level scaffolds
· Deterministic outputs
· Prompt versioning
· Eval pipelines

Cloud & DevOps
· GCP: Cloud Run · Cloud Build
· Artifact Registry · IAM
· Docker · Kubernetes · Terraform
· CI/CD · Blue-green · Rollback
· LangSmith · Helicone observability

Engineering Stack
· Python · Next.js · React · TypeScript
· Node.js · FastAPI · Redis
· PostgreSQL · Supabase · Stripe
· LangChain · LlamaIndex · Vercel

Architecture
· Multi-tenant SaaS · Row-Level Security
· API Gateway · Event-Driven
· Semantic Caching · Zero-Trust RBAC
· ADA Compliance · Security Audits
· Machine Learning · Fine-tuning · RLHF

AI Models & Tools
· Claude Sonnet/Opus Thinking · Haiku 4.5
· GPT-4o · GPT-OSS 120B · o1/o3
· Gemini 3.1 Pro · Flash · Llama 3
· DeepSeek R1 · Mistral · Kimi · Perplexity
· Claude Code · ChatGPT Codex · Manus · Nano Banana · Cursor

Visual Design & Media
· UI/UX Design · Figma · Component Systems · Design Tokens
· Brand Identity · Visual Direction · Creative Direction
· Photography (Studio/Fashion/Editorial/Aerial)
· Video Production · Scriptwriting · Editing
· ElevenLabs · Veo · Sora · Kling · Seedance

FEATURED PROJECTS

CafeOS

cafeos.co

Zero-Commission Restaurant OS — 20+ Paying Clients — Built in 45 Days

Agentic AI · Multi-tenant · GCP · CI/CD · Stripe · RBAC · ADA

12-module enterprise restaurant platform built in 45 days via agentic AI orchestration. Features GenieUs AI agent (SMS/email/WhatsApp updates), TV Kiosk Ordering (industry first), Pulse Analytics, Magic Menu, QR Dine-In, and Stripe payments. 20+ paying restaurants in TX, CA, IL. $100–$350/month per client. Targeting 1,000 restaurants and $225K MRR by 2027.

Google Antigravity Prompt

kasaju.com

Single-Prompt Full-Stack SaaS Scaffold

Prompt Engineering · Claude · Next.js · Deterministic · 10× Velocity

Master system-level prompt that scaffolds a complete production Next.js SaaS app — auth, RBAC, analytics, database schema, and all modules — from a single prompt with zero manual configuration. Powers CafeOS, KickHub, and SalonOS build cycles. Demonstrates 10× developer velocity.

KickHub

kickhub.us

AI-Native Soccer League Management Platform

Domain AI · Sports Tech · GotSport Competitor · Next.js

Full-stack soccer league management: scheduling, registrations, standings, coaching tools, and communications. Competes directly with GotSport in the US youth sports market. Built AI-first validating the reusable vertical SaaS architecture across competitive sports management.

SalonOS

AI-Native Salon & Spa Operations Platform · In Active Development

Booking · POS · Staff Scheduling · Multi-tenant · AI-first

Full-stack salon operations platform currently in active development: online booking, POS, staff scheduling, inventory management, client CRM, and loyalty programs. Applies the same agentic AI build methodology as CafeOS — validating the reusable vertical SaaS pattern across hospitality and service verticals beyond food & beverage.

AI ENGINEERING APPROACH

LLM Selection & Cross-Model Validation

Route by task — Claude Opus/Thinking for reasoning, Haiku for throughput, GPT-OSS 120B for cost, Gemini for multimodal, DeepSeek for technical depth. Model strengths shift daily; selection is updated continuously. Critical outputs are cross-validated across models — no single LLM is trusted unconditionally in production.

Multi-Agent System Design

Orchestrator-worker hierarchies with persistent state, tool registries, and human-in-the-loop escalation. Deployed in CafeOS across operator, customer, and analytics agent layers.

RAG & Production Evaluation

Chunking strategies (semantic/recursive/fixed), hybrid retrieval, cross-encoder re-ranking. Quality tracked via RAGAS. p95 latency and TTFT monitored per endpoint in production.

GCP · CI/CD · Production Safety

Cloud Run + Cloud Build + Artifact Registry. Blue-green deployments with automated rollback. Output validators, hallucination monitoring (TruLens/RAGAS), ADA compliance, and security audits.

UI/UX · Design Systems · Product Craft

All platforms designed from scratch in Figma — component libraries, design tokens, and responsive systems. Deep product design background ensures consumer-quality experience at enterprise scale.

IMPACT & METRICS

20+	$225K	45	1,000+	$4,200	10×
Paying Restaurant Clients	Projected MRR by 2027	Days CafeOS Built End-to-End	Restaurant Target 2027	Monthly Savings Per Restaurant	Developer Velocity Gain

EDUCATION & PROFESSIONAL DEVELOPMENT

Bachelor of Media Technology (BMT) · 2007

Shepherd College · Purbanchal University · Nepal

Computer Science, Advertising & Marketing, Broadcasting, Web Design, Editorial Publishing, Reporting, and Design.

Foundation in Computer Science · NCC Education UK / Cambridge · 1996–1997

Programming (PASCAL), data structures, algorithms, data networking, and systems design.

Continuous Learning — AI Engineering · 2021–Present

· Hundreds of hours from leading AI researchers and engineers
· Deeplearning.ai · Fast.ai · Anthropic · Google AI · OpenAI
· Production-validated through live, revenue-generating systems

Skills demonstrated through production output, not certificates.

Languages: English (Fluent) · Nepali (Native)