

PRADEEP REDDY

Founder & CTO | Solution Architect

Los Angeles, CA | pradeepreddi@gmail.com | [linkedin.com/in/pradeepsreddy](https://www.linkedin.com/in/pradeepsreddy)

PROFESSIONAL SUMMARY

Solution Architect and Senior Software Engineer with 18+ years architecting, building, and securing enterprise systems. Now applying that depth to production-grade AI platforms and AI-native software engineering methodology as Founder & CTO of RAYFINITE LLC. Designed and operate a tiered, multi-agent AI engineering system (Opus/Sonnet/Haiku-routed, role-specialized, human-gated) and use it to architect and build a twelve-venture technology portfolio end to end, including the RAYFINITE Agentic Framework, the multi-agent engineering platform the whole portfolio is built through, and the RAYFINITE web and intranet platform that runs the company: identity and access control, HR, goals and approvals, asset tracking, and an append-only audit trail. Hands-on across the modern AI stack: Anthropic Claude API, Apple Foundation Models, Google Gemini and Vertex AI, MCP servers, LangChain, RAG, multi-agent orchestration, and edge inference (llama.cpp, GGUF, MLC-LLM, Ollama). Combines this AI practice with deep cloud (AWS, Firebase/GCP, Oracle Cloud), security (NIST 800-53), and full-stack enterprise experience, including leading a team of three at a Maryland state agency.

TECHNICAL SKILLS

Agentic AI Engineering: Multi-Agent Orchestration (Opus/Sonnet/Haiku Tiered Routing), Role-Specialized Agents, Adversarial Verification, Parallel Fan-Out + Synthesis, Human-Gated Deploys, Persistent Cross-Session Memory, Skill Supply-Chain Allowlisting

AI Architecture: Retrieval-Augmented Generation (RAG), Multimodal Pipelines (Vision, Voice, Text, GPS), On-Device / Edge Inference, Hybrid Cloud-Edge AI, MCP Servers (author), Safety-by-Design, Confidence Scoring, Threat Classification, Dual-Layer Guardrails

AI Models & Inference: Anthropic Claude (Opus/Sonnet/Haiku), Apple Foundation Models, OpenAI GPT-4, Google Gemini (Gemini Live API), Google AI Studio, Google Vertex AI, Cloudflare Workers AI, Groq, NotebookLM, LLaMA 3, LLaVA, Whisper, Piper, llama.cpp, MLC-LLM, GGUF Quantization, Ollama

AI Tooling & Frameworks: Model Context Protocol (MCP) Servers, LangChain, Tool Use / Function Calling, FAISS, pgvector, SQLite-VSS, MiniLM Embeddings, LoRA Adapters, Fine-Tuning Pipelines, Argilla, Multi-Provider AI Routing + Circuit Breaker

Agentic Coding: Claude Code, Cursor, Google Antigravity, GitHub Copilot, VS Code, Pair-Programming with LLMs

Cloud & DevOps: AWS (EC2, S3, Lambda, Route 53, ELB, VPC, VPN, IAM), Firebase (Cloud Functions, Firestore, App Check, Secret Manager), Google Cloud Platform, Oracle Cloud (proof-of-concept), VMware, GoAnywhere MFT, Docker, Kubernetes, GitHub Actions, CI/CD, Infrastructure as Code

Application Stack: iOS (Swift 6, SwiftUI, StoreKit 2), React, React Native, Next.js, Tauri (Rust), Node.js, TypeScript, Python, FastAPI, WebSockets, WebFlow, GovDelivery

Programming Languages: JavaScript, HTML5, CSS, C#, .NET, VB.NET, Java, TypeScript, Python, Swift

Operating Systems: Unix, Ubuntu, Linux, Android OS, macOS, Windows OS

Data & Storage: Firestore, PostgreSQL, pgvector, SQLite, Redis, Microsoft SQL Server, MongoDB, Oracle 12c, Oracle AI Database, DB2, Power BI, Oracle WebCenter, OBIEE

Security & Compliance: NIST 800-53, Pre-Deploy Security Audits, Secret Manager (no secrets in source), SHA-256 API Key Hashing, Firestore Security Rules, Zero-Trust Credentials, RBAC, Tenant Isolation, Audit Logging, Palo Alto Firewalls, Splunk, Tenable, Security Onion, Kali Linux, Shodan, Oracle Wallet, SIEM

Architecture & Leadership: Solution Architecture, Technical Architecture Documents (TADs), ADRs, Cloud-Native Design, Microservices, API Integration, Unit Economics, Team Supervision, Mentoring, SDLC Ownership

Languages: English (professional), Telugu (native), Hindi (fluent)

PROFESSIONAL EXPERIENCE

RAYFINITE LLC, Los Angeles, CA

Jan 2026 - Present

Founder & CTO | Solution Architect

Founded a solo venture studio building AI-native products end-to-end (architecture, engineering, security, and launch) operated through a proprietary multi-agent AI engineering platform. Key ventures:

- **RAYFINITE Agentic Framework:** Built a multi-agent AI software-engineering platform: tiered model routing, role-specialized agents, fail-closed CI security-review gates verified adversarially with bait pull requests, cross-machine agent federation, and skill supply-chain allowlisting; packaged as a versioned installable plugin.
- **RAYFINITE Web & Intranet:** Built and operate the RAYFINITE public site and the internal platform behind it: identity and access control, HR, goals and approvals, asset tracking, and an append-only audit trail. Both are live and still growing.
- **WeatherHappiness:** Designed and built an iOS wellbeing app (SwiftUI + HealthKit) correlating weather patterns with personal health metrics, with a companion web experience on Firebase/GCP.
- **Suncalp | HonestResolve:** Architected an AI-guided personal-accountability app across iOS, Android, and web: commitment-authenticity scoring, streak-based check-ins, and social accountability features on a Firebase-backed architecture.
- **LovelyRobo:** Architected a spatial AI-companion platform: photoreal 3D companion robots as task-specialists with live voice conversation and persistent memory, offered individually and as a bundle.
- **Regionara:** Architected a multi-tenant community platform serving several separately branded diaspora communities from one codebase: local business directory, cultural events, jobs, and forums, with tenant isolation and a retrieval-augmented search layer running embeddings locally (Next.js, React, Postgres with pgvector).
- **Zenolon:** Architected an ERP intelligence platform: four core modules (Governance/Identity, HRMS/HCM, Financials, Zenolon Engine) that other RAYFINITE products integrate with as a shared integration layer.
- **FactGuard:** Delivered an evidence verification engine on Firebase/GCP, live across web, a public REST API, and an MCP server, with a proprietary multi-step verification engine, multi-provider model routing with a circuit breaker, and live Stripe subscription billing. The iOS app is built and preparing for launch.
- **Dataveb:** Architected a no-code web-scraping and automated-reporting SaaS (Python/FastAPI, Next.js, Playwright), security-first: SSRF-guarded extraction, idempotent job scheduler, Stripe billing with per-tenant usage metering.
- **DiamondHull:** Designed and built a portable purple-team platform for home and small-office networks: passive-first intrusion detection that never sits in the path of a legitimate device, decoys and honeypots, verifiable evidence capture signed through a swappable asymmetric signing port specified for Ed25519, and manual-by-default response so a false positive never cuts off a real device.
- **AEGIDON:** Architected a cloud/edge hybrid multi-agent platform for IT incident response: incident-commander orchestration across six specialist agents, zero-trust credential handling, blast-radius analysis, and end-to-end audit logging (FastAPI, WebSockets, PostgreSQL, Redis).
- **AThEnA:** Architected a decision intelligence platform for regulated and sovereign environments, integrating model-agnostic inference, mission alignment, risk analysis, decision workflow, and audit into one system so a decision can be traced, explained, and defended rather than merely produced. Fourteen functional domains across Corporate and Government editions, designed to deploy across cloud, on-premises, and classified environments, with requirements locked through BRD, PRD, a twenty-section FRD, a design-patterns catalog, and a conceptual model.

Maryland Longitudinal Data System Center (MLDSC), Baltimore, MD

Jan 2015 - Jun 2026

IT Technical Specialist Supervisor | Solution Architect

- Led the agency's digital footprint end to end: solution architecture, development, and operations for a public-facing site serving 100,000+ annual visitors, an intranet portal, and a stack of integrated enterprise applications (GoAnywhere MFT, GovDelivery, Power BI, Oracle WebCenter / OBIEE, WebFlow, Tenable, Palo Alto, and more).
- Prototyped a Retrieval-Augmented Generation (RAG) solution on Google Vertex AI, Google AI Studio, and NotebookLM to surface insights from internal agency documents, establishing a reusable pattern for LLM-powered knowledge management.
- Evaluated locally-hosted open-source LLMs via Ollama and LLaMA 3, integrated through LangChain and Model Context Protocol (MCP) servers, for secure on-premise AI under data-sovereignty constraints.
- Drove a 2-3x acceleration in team delivery velocity by adopting AI-assisted, agentic coding workflows with Claude Code, Cursor, GitHub Copilot, and VS Code.
- Architected and led migration of the agency's infrastructure to the Tierpoint datacenter, 10 physical servers hosting a VMware-based estate of ~40 virtual servers and 30 desktop VMs, delivering a 50%+ improvement in system performance and decommissioning legacy platforms.

- Spearheaded AWS cloud adoption and implemented the agency's AWS infrastructure: disaster recovery, application hosting, and off-site storage via EC2, S3, Route 53, Lambda, ELB, VPN, and VPC networking, and evaluated Oracle Cloud through proof-of-concept work.
- Modernized legacy Oracle WebCenter executive dashboards by migrating analytics workloads to Microsoft Power BI, improving refresh times and lowering Oracle licensing exposure.
- Strengthened agency cybersecurity posture by developing NIST-aligned firewall and detection rules, partnering with the SOC team on Palo Alto Firewalls, and using Splunk and Tenable for SIEM event analysis and vulnerability management.
- Supervised and mentored a team of three IT staff, partnering with the agency's governing board to translate strategic priorities into technical roadmaps, and owned the full SDLC.

Tejase Technologies Inc., Des Moines, IA

Aug 2011 - Jan 2015

Senior Software Engineer | Technical Consultant

- **Logitech Inc. (Newark, CA), Project Manager:** Delivered enterprise architecture and technical consulting across Fortune 500 and public-sector clients, leading complex initiatives from solution design through production rollout.
- **Catamaran Rx (Rockville, MD; now OptumRx), Senior Application Architect / ECM Consultant:** Architected and implemented enterprise content management (ECM) solutions using Oracle Fusion Middleware and modern web technologies, improving document workflow efficiency for a healthcare benefits leader.
- **Minnesota Housing Finance Agency (St. Paul, MN), Senior Technical Consultant:** Drove design and execution of strategic IT modernization initiatives: web application development, integrations, and platform stability for a state housing finance agency.

Fusion Global Solutions, Overland Park, KS

Feb 2008 - Aug 2011

Software Engineer | Technical Consultant

- Delivered application development and technical consulting using ASP.NET, C#, JavaScript, jQuery, and SQL Server across clients including Minnesota Housing, the Commonwealth of Massachusetts (Healthcare and Financial Policy), AmeriHealth Administration, the State of Nebraska (DHHS), Kaplan Eduneering, Hearst Corporation, and the Florida Senate (Technical Consultant / Senior Software Engineer: Microsoft .NET web services and SQL Server).

EDUCATION

Master of Science, Computer Science

Jan 2006 - May 2008

New York Institute of Technology, Manhattan Campus, NY

Specialization: Artificial Intelligence, Operating System Security, Compiler Theory, Software Engineering, Computer Graphics.

Bachelor of Technology, Computer Science & Information Technology

Jun 2001 - May 2005

Jawaharlal Nehru Technological University (JNTU), Hyderabad, India

Coursework: Data Structures, Operating Systems, Database Management, Information Security, Cloud Computing.

CERTIFICATIONS

- Microsoft Certified Professional

PROFESSIONAL DEVELOPMENT

- **Anthropic:** Multiple AI / LLM courses with completion certificates: prompt engineering, Claude API integration, applied AI workflows.
- **AWS Training & Certification (Instructor-Led):** Architecting on AWS (3 days), Developing on AWS (3 days), AWS Cloud Practitioner Essentials Day.
- **State of Maryland:** Multiple cybersecurity training programs: security awareness, NIST controls, incident response.
- **Self-Directed AI Engineering:** Hands-on architecture and engineering work with Anthropic Claude API, Apple Foundation Models, Firebase/GCP, Vertex AI, Google AI Studio, NotebookLM, LangChain, MCP servers, Groq, Google Antigravity, Ollama, LLaMA 3, llama.cpp, and Jupyter Notebooks.

INTERESTS

Filmmaking · Film · Audiobooks · Reading · Exploring