BODA MADHUKAR REDDY

AI Architect & Senior .NET Engineer | Agentic AI | Self-Hosted LLMs

+91 95731 53479 madhukarreddyboda538@gmail.com
linkedin.com/in/madhukar-reddy-boda dhucar.in github.com/Madhukar538
Hyderabad, India



WORK EXPERIENCE

Senior Software Engineer (R&D)

Revalsys Technologies

Dec 2021 – Present Hyderabad, IN (Full-time)

AI platforms

- Runs a team of AI coding agents in Herdr: Claude Code leads and plans, Antigravity builds, Copilot and Codex review, guided by written specs, agent skills and multi-agent review, with every change owned and merged by hand.
- Architected a multi-tenant RAG chatbot platform on .NET 10 (Microsoft Agent Framework agents, pgvector, text-to-SQL, MCP tools, PII guardrails and SignalR streaming), then led an end-to-end latency audit that cut typical answer time from 30–40 s to about 13 s.
- Built an MCP (Model Context Protocol) server exposing 40+ tools that give AI assistants a map of C# codebases and SQL Server schemas, using Roslyn code analysis and schema extraction synced to MongoDB.
- Designed a hybrid product search engine on Solr 9 that fuses BM25 keyword and kNN vector results with Reciprocal Rank Fusion.
- Moved AI workloads to self-hosted LLMs (Ollama on an in-house GPU server), saving about ₹10,000 a month in cloud AI costs.
- Integrated AI into internal workflows, including NLP chatbots, RAG systems and automated code review.

Performance & tooling

- Created repeatable k6 load-testing suites with Grafana dashboards, setting SLA baselines and capacity models for mission-critical APIs.
- Diagnosed full collection scans behind slow conversation queries and fixed them with compound and unique MongoDB indexes, cutting query latency from 500–2,000 ms to 5–50 ms.
- Developed a CI/CD code-review assistant on local LLMs, reducing PR review cycle time by 30%.
- Engineered enterprise data pipelines and SAP integration channels with zero-loss transactional sync.

Architecture & R&D

- Prototyped an offline-first procurement and inspection app (Next.js PWA with Android SQLite) with outbox-based sync (in progress).
- Extended Uptime Kuma into an uptime and incident-management platform with a .NET API, push alerts and ticketing (in progress).
- Led all of the company's .NET R&D projects as the sole lead, from rapid prototypes to working systems.
- Migrated RevalERP's legacy systems to modern .NET hybrid architectures.
- Produced proofs of concept for IoT (MQTT), debugging tools and context-aware assistants.
- Implemented an RDLC reporting application for internal workflows, from requirements through delivery.

Client delivery

- Works directly with the clients who run Revalsys products: understands what they need, scopes it into releases and ships it into their production.
- Delivered core features for 5 high-traffic e-commerce brands (Jockey, Speedo, Manyavar, LuxCozi and Fossil India) on Angular and .NET.

SUMMARY

AI architect, agentic full stack developer and senior .NET engineer with nearly 5 years of building and scaling high-traffic .NET systems and self-hosted AI platforms. Works forward-deployed: sits with the clients who run Revalsys products, scopes what they need and ships it into their production. Runs a team of AI coding agents (Claude Code leads and plans, Antigravity builds, Copilot and Codex review) with specs, skills and review, and owns what they ship: a product foundation in two days, and legacy worker services moved to .NET 10 in two. Cut a multi-tenant RAG chatbot's answer time from 30–40 s to about 13 s and MongoDB query latency from up to 2 s to under 50 ms. Works across .NET 10, Microsoft Agent Framework, MCP, Solr 9 hybrid search and k6 + Grafana.

SKILLS

Languages

C# TypeScript JavaScript Python SQL C

Agentic Development

Claude Code OpenAI Codex
GitHub Copilot CLI Google Antigravity Cursor
Herdr Agent Skills MCP servers
Multi-agent review AI evals

Backend

.NET 10 / ASP.NET Core Web API & MVC
Background & Worker Services Blazor Node.js
FastAPI

Frontend & Mobile

Next.js 16 (App Router, ISR) React 19 Angular
Tailwind CSS React Native Offline-first PWA

AI & LLMs

RAG Semantic Kernel
Microsoft Agent Framework
MCP (Model Context Protocol) Ollama Qwen3
Gemma LiteLLM mem0 LangChain
ONNX Runtime Gemini VLM Document AI
WebLLM whisper.cpp Piper OpenCV

- In early 2026, migrated Fossil India from Angular to Next.js and from .NET Core 3.1 to .NET 10 with a team of AI coding agents, and shipped it to production.
- Led the ONDC (Open Network for Digital Commerce) protocol integration.
- Integrated payment gateways (Juspay, Paytm and PayU), GoKwik checkout and Shiprocket shipping into the e-commerce platforms.
- Shipped core modules across the 6-product RevalSuite: RevalERP, RevalHRM, RevalCRM, RevalPOS, RevalCMS and RevalInventory.
- Built Reval Meet, a video conferencing platform on WebRTC, SignalR and Node.js.

PROJECTS

Multi-Tenant AI Chatbot Platform

Embeddable RAG assistant on .NET 10: Microsoft Agent Framework agents with MCP tools for database and docs access, pgvector retrieval, interactive map answers, passkey sign-in, server-bound licensing and remote installs and upgrades.

Result: Most answers now arrive in about 13 seconds, down from 30–40, after a latency audit and a move to Microsoft Agent Framework.

Technologies: .NET 10, Microsoft Agent Framework, MCP, pgvector, Next.js 16

Offline-First Procurement & Inspection App

Next.js PWA plus Android shell for field procurement: one storage interface over Android SQLite and IndexedDB, master-data sync, and a transactional outbox that replays actions when back online.

Technologies: Next.js, .NET, SQLite, IndexedDB, Android

Uptime & Incident Management Platform

Uptime Kuma as a headless check engine behind a custom .NET 10 API and React admin: incidents, maintenance windows, FCM push alerts, flapping suppression, self-healing hooks and ticketing.

Technologies: .NET 10, Uptime Kuma, React Router, PostgreSQL, Docker

dhucar.in: Portfolio Platform

This site, built and run end to end: Next.js 16 pages kept fresh by on-demand revalidation from a .NET 8 API, self-hosted MongoDB, passkey sign-in for the admin, live presence over SignalR, a read-only MCP endpoint for AI assistants and eight switchable designs, deployed with Coolify behind a Cloudflare Tunnel.

Technologies: Next.js 16, .NET 8, MongoDB, SignalR, MCP, Coolify

RESEARCH

Local LLM GPU Benchmark: RTX 5090 vs DGX Spark

Same model (Gemma 4 26B), same Ollama version, same agent request: ~15 s on an RTX 5090 against ~41 s on a DGX Spark. Memory bandwidth (1,792 vs ~273 GB/s) mattered more than memory size once the model fit.

Technologies: Ollama, Gemma 4, RTX 5090, DGX Spark

Hybrid Product Search (BM25 + Vector + RRF)

Solr 9 engine fusing keyword and kNN vector results with Reciprocal Rank Fusion. In-process ONNX embeddings, rule-based NLU over a live catalog vocabulary, typo tolerance, price parsing and multi-item queries.

Technologies: Solr 9, ONNX Runtime, MiniLM, RRF, .NET

Search & Vectors

Solr 9 (kNN, Suggester, MLT) · Hybrid BM25 + Vector (RRF) · pgvector · Qdrant · Embeddings

Data & Caching

SQL Server · PostgreSQL · MongoDB · ClickHouse · SQLite · IndexedDB · Redis · Memcached · EF Core · Dapper · ADO.NET

Messaging & Realtime

RabbitMQ · SignalR · gRPC · MQTT · WebRTC · Socket.IO · Firebase (FCM)

Payments & Commerce

Juspay · Paytm · PayU · GoKwik checkout · Shiprocket shipping · ONDC

DevOps & Infrastructure

Docker · Docker Swarm · K3s · Proxmox · Coolify · Cloudflare Tunnel · Terraform · Azure · Azure DevOps / TFS · IIS

Observability & Quality

k6 · Grafana · Uptime Kuma · Load Testing & SLAs · Playwright · Roslyn Analyzers · Postman

EDUCATION

B.Sc in Computer Science

Satavahana University

Karimnagar, Telangana