

Naveen Hemanth Kokkonda

+91-6300296968 | naveenhemanth4@gmail.com | linkedin.com/in/hemanthkokkonda | github.com/hemanth090 | x.com/HemanthKokkonda

EDUCATION

Sreyas Institute of Engineering and Technology

B.Tech, Computer Science and Engineering (Data Science) - CGPA: 7.51/10

Hyderabad, India

Oct. 2022 – Jun. 2026

EXPERIENCE

AI Engineer (Full Stack)

Dec. 2025 – Aug. 2026

Tradl AI - Stock Research and Trading Platform

On-site, Bangalore

- Cut initial render time 40% and heap usage 60% on financial data tables by virtualizing 1,000+ live stock tickers with TanStack Virtual, then extended the approach to 2 more table surfaces (Discovery, Playground)
- Delivered real-time pattern-detection visualization by deploying an AWS-hosted, WebSocket-fed frontend that streams live ticker bars and LTP (last traded price) directly onto TradingView charts
- Enabled citation-grounded Q&A over 45 financial PDF filings with a hybrid retrieval pipeline combining Qwen3-Embedding-4B dense vectors with sparse search, followed by Qwen3-Reranker-4B cross-encoder reranking
- Orchestrated an agentic workflow spanning 3 task types (document retrieval, structured financial-metric lookup, multi-step reasoning) by dynamically routing across 3 LLMs (Claude Sonnet, Claude Haiku, and Bedrock Qwen3-235B) based on task complexity
- Lifted retrieval recall 65% (0.37 to 0.61) on a 90-question financial QA benchmark by designing ingestion and indexing pipelines with Qwen3-VL page-level extraction for charts and tables, processing 45 financial filings into 30,137 indexed points

PROJECTS

Articulate - arXiv Research Summarizer | [Live Demo](#) | [GitHub](#)

Sep. 2026

Technologies: Next.js 15, TypeScript, Groq, Vercel AI SDK, arXiv API, OpenRouter

- Built a research summarizer turning a topic query into structured summaries from the top 20 arXiv papers via the official arXiv API (Atom XML parsing) with rate-limited caching across 2 Next.js API routes.
- Re-ranked all 20 papers in one call with TypeSafe Jev-1.13 via OpenRouter Decisions API using 20 parallel Score questions, sorting by score $\times$ confidence with per-paper scores and graceful fallback to arXiv order when offline.
- Streamed crisp summaries (Overview, Key Ideas, Method, Results, Why It Matters) with Vercel AI SDK on Groq gpt-oss-120b using best-effort full text (4s timeout) with abstract fallback, deployed on Vercel with server-side-only keys.

Scout - Autonomous Research Agent | [GitHub](#)

Aug. 2026

Technologies: Python, FastAPI, LangGraph, React, Supabase, Redis

- Developed an autonomous research platform using Python, FastAPI, LangGraph, and React that converts natural-language queries into structured, citation-backed reports using arXiv and live web sources.
- Designed a multi-agent LangGraph workflow with a supervisor orchestrating parallel research, writer-critic revision, and guard agents using isolated per-subgraph state and Send fan-out/fan-in for modularity and fault isolation.
- Shipped safeguards with human-in-the-loop approval below 0.70 confidence, citation verification, PII redaction, 3 retries before fallback, and hard budgets of \$0.50/run and \$5.00/day; monitored via SSE streaming, Supabase, Redis, and a React dashboard.

TOON Studio - Token Efficiency Toolkit | [Live Demo](#) | [GitHub](#)

Dec. 2025

Technologies: Python, FastAPI, React, Gemini, LangSmith, Vite

- Launched a full-stack JSON/TOON converter (FastAPI, React/Vite) measuring about 33% token reduction on simple objects and 52% on large arrays via its Gemini-tokenizer harness, plus a live cost dashboard (\$0.15/1M in, \$0.60/1M out).
- Integrated Google Gemini 2.5 Flash with LangSmith tracing and WebSocket streaming for real-time AI-powered analysis with live token-usage monitoring, deployed as a live demo on Vercel.
- Shipped reliability with a 10-test pytest suite covering conversion and token counting plus request/response logging across the FastAPI-to-React pipeline, easing debugging of conversion edge cases.

TECHNICAL SKILLS

Languages & Frameworks: Python, TypeScript, SQL, React, FastAPI, Tailwind CSS

AI Systems: Hybrid Retrieval + Reranking, RAG Ingestion/Indexing, Agentic Workflows (LangGraph), LLM Streaming (Vercel AI SDK), Tracing/Eval (LangSmith)

Cloud & Data: PostgreSQL (Supabase/pgvector), MongoDB, Redis, FAISS/Qdrant, AWS Bedrock/EC2, Vercel

CERTIFICATIONS

- Oracle Cloud Infrastructure 2025 Certified Generative AI Professional (Aug 2025) [[View Certificate](#)]
- Completed Simplilearn Front End Development covering HTML, CSS, JavaScript (Jun 2025) [[View Certificate](#)]
- Certified in Core Python Programming by Infosys Springboard (Sep 2024) [[View Certificate](#)]