

Naveen Hemanth Kokkonda

+91-6300296968 | naveenhemanth4@gmail.com | hemanthk.xyz | 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

- Built citation-grounded Q&A over 45 financial filings (30,137 chunks) with dense + sparse retrieval and a Qwen3 cross-encoder reranker, lifting recall 0.37 → 0.61 (+24 pts) on a 90-question evaluation set.
- Extended ingestion with page-level Qwen3-VL extraction after the first parser flattened charts and tables into broken text, making those answers citable.
- Benchmarked embedding, reranker, OCR/VLM and quantized candidates plus agent SDKs (Anthropic, OpenAI, Chinese open-source) on accuracy, p95 latency and cost; cached embeddings and parses.
- Routed queries by task type across Claude Sonnet, Claude Haiku and Bedrock Qwen3-235B, keeping routine work off the largest model.
- Instrumented the RAG and market services with structured logging, metrics and CloudWatch alarms, so failures surface in dashboards instead of user reports.
- Virtualized 1,000+ live tickers with TanStack Virtual (initial render -40%, heap -60%, reused on Discovery and Playground) and streamed WebSocket ticks onto TradingView charts.
- Designed the stock screener's agent-status UI: rotating status copy over a CSS shimmer, driven by live tool calls and state transitions, so long scans read as progress, not a frozen screen.

PROJECTS

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

Sep. 2026

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

- Shipped an arXiv summarizer: topic query to structured summaries of the top 20 papers via the official API (Atom parsing, rate-limited cache in 2 Next.js routes), streamed in five sections on Groq gpt-oss-120b.
- Ranked the 20 papers with a structured scorer (Jev 1.13 via OpenRouter: 20 parallel rubric scores in one call) instead of arXiv order, with a 4s full-text timeout that falls back to abstracts.

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

Aug. 2026

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

- Orchestrated a LangGraph pipeline (supervisor, parallel researchers, writer-critic, guard) that turns a query into a citation-backed report over arXiv and live web, hard-capped at \$0.50/run and \$5.00/day.
- Hardened the pipeline with citation verification via Claude web search, PII redaction (email, phone, card), 3 retries and human approval below 0.70 confidence; wrote an LLM-as-judge harness scoring reports 0 to 2 per required fact.

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

Dec. 2025

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

- Launched a JSON/TOON converter with a tiktoken (cl100k_base) harness that cut tokens 29% vs compact JSON and 58% vs pretty-printed JSON (median over 20 payloads), plus a live cost dashboard.
- Integrated Gemini 2.5 Flash with LangSmith tracing and WebSocket streaming for live token-usage analysis.

TECHNICAL SKILLS

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

AI Systems: Hybrid retrieval + reranking, RAG ingestion/indexing, agentic workflows (LangGraph), LLM streaming (Vercel AI SDK), structured outputs

Evaluation & Reliability: hand-labeled QA sets, recall@k, citation verification, LLM-as-judge rubric scoring, tracing (LangSmith), cost and latency budgets

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]