

Venkata Sai Kumar Vendoti

AI Engineer | Agentic AI · LLM Systems · RAG

Hyderabad, Telangana • venkatasaikumarvendoti@gmail.com • +91 7995451956 • [linkedin.com/in/venkata-sai-kumar-vendoti](https://www.linkedin.com/in/venkata-sai-kumar-vendoti)

PROFESSIONAL SUMMARY

AI Engineer with 3 years building **production agentic AI and LLM systems end-to-end** — tool-calling agents, MCP, RAG/embeddings, and multi-model GenAI — backed by strong full-stack and cloud engineering (Python, TypeScript, Next.js, GCP, AWS, Kubernetes). Track record of shipping applied AI with measurable cost and performance wins.

EXPERIENCE

Fn7 — AI Engineer (Intern → Full-time)

Hyderabad • Jul 2023 – Present

AGENTIC AI & LLM SYSTEMS

- Built the **Brain agent end-to-end** — a **25-tool** brand-memory system that builds a knowledge graph (company, products, personas, competitors, assets) so every downstream AI (content, campaigns, SEO) operates in consistent brand voice; rebuilt on **Google ADK** and running in production.
- Architected the **Agent Builder** (full frontend + backend) — a no-code platform defining agents as **pathways** → **stages** with pre/post server-side callbacks, auto-generated prompts, and autonomous execution, including a meta-agent that generates other agents' definitions and prompts.
- Developed a custom **Model Context Protocol (MCP) server** exposing LLM tool-calling across **12 platform APIs** (LinkedIn, X, Reddit, Instagram, Apollo, internal knowledge graph) so agents can post, comment, and pull live brand context on the user's behalf.
- Engineered a **6-provider LLM router** (OpenAI, Google Vertex/Gemini, Anthropic, Azure/Grok, AWS Bedrock) with cross-provider fallback chains, per-call token/cost accounting, and an **LLM-as-judge** quality gate.

RAG & EMBEDDINGS

- Designed and built a **hybrid RAG pipeline** for SEO keyword-coverage analysis — dense **BGE embeddings + BM25, reciprocal-rank fusion**, and **cross-encoder reranking**, plus a score-calibration framework mapping model scores to a P(covered) coverage decision.
- Cut AI audit cost $\approx 40\times$ (≈ 120 LLM calls → 4) by batching grounded probes, and reduced latency/memory with embedding-model singletons and content-hash scoring caches — with no loss of per-question citation accuracy.
- Prototyped a **competitive price-intelligence pilot for HP** — used embeddings to semantically match the same and comparable laptops across **Amazon, Walmart, and Flipkart**, benchmarking vendor and competitor (Dell, Intel, etc.) pricing by segment to inform HP's own pricing.

GENAI PRODUCT

- Built the **Video Studio generation pipeline end-to-end** — avatar + voice videos on **Google Veo 3.1** and **ElevenLabs** voice cloning, with **character-consistent multi-clip continuation** (last-frame chaining) and an AssemblyAI + FFmpeg transcription/caption pipeline.
- Shipped **multi-platform ad-campaign launch** across **Google, Meta, LinkedIn, and Reddit** — parallel fan-out with partial-failure handling, **per-platform rollback**, and an LLM-assisted error classifier — consolidating cross-platform performance into one interface.
- Created the **MVP Factory** — an agent that turns a founder's idea into a working mock product, generating user roles, flows, and shared screens rendered through a **low-code JSON-Forms engine** (AI fills component type + data, no code-gen), with a full inline edit mode.

FULL-STACK, PERFORMANCE & LEADERSHIP

- Cut page-load time from **5–6s to sub-second** via bundle code-splitting, landing-first load with background prefetch, Firebase index caching, **TanStack Query**, and entity-based cache invalidation.
- Led migration from **Angular to Next.js** (React 19, Module Federation micro-frontends); built Firebase authentication and sign-up/login flows.
- Owned CI/CD and infrastructure — **Docker, Terraform, Kubernetes** (ingress, scaling) — and migrated workloads from **Azure AKS to AWS ECS**.
- **Managed a 3-member intern team** on client projects — ran client calls, translated feedback into tasks, and distributed work.

INTERNSHIP HIGHLIGHTS (JUL 2023 – 2024)

- Delivered a **veterinary AI platform solo** — extracted pet medical history from documents (**Google Cloud Vision**) and treatment notes from voice (**Speech-to-Text**) to auto-generate visit summaries and discharge reports; fixed a cross-platform (Windows/Mac Chrome) audio bug.
- Engineered custom **JSON-Forms (Helica) renderers** (grid/flex/array layouts, cards, headings) and fixed a form-instance data-sync issue enabling live cross-page updates.

TECHNICAL SKILLS

AI / LLM: Agentic systems, MCP (tool-calling), RAG, embeddings, hybrid retrieval (dense + BM25), reranking, prompt engineering, LLM-as-judge, multi-model orchestration, applied machine learning, Google ADK

Models & APIs: OpenAI (GPT-4o/5), Google Gemini + Veo (Vertex AI), Anthropic Claude, xAI Grok, AWS Bedrock, ElevenLabs (TTS/voice), Fal.ai

Languages: Python, TypeScript, JavaScript

Backend: Node.js, FastAPI, Express, REST, Firebase (Firestore/Auth/RTDB), Elasticsearch

Frontend: React, Next.js, Angular, Module Federation, TanStack Query, Tailwind

Cloud / DevOps: GCP (Vertex AI), AWS (ECS/Bedrock), Docker, Kubernetes, Terraform, CI/CD, Fastly CDN

Practices: Agile, cross-functional collaboration, client-facing delivery, mentoring

EDUCATION

B.Tech, Computer Science — RGUKT RK Valley | 2020 – 2024 | CGPA: 8.77/10