

SESHANAGOTTU VENKATA SUJITH PRAVEEN

AI/ML ENGINEER • AGENTIC SYSTEMS & RETRIEVAL-AUGMENTED GENERATION

Mumbai, India • +91 8332816876 • svspraveens@gmail.com • svspraveen.vercel.app • linkedin.com/in/svs-praveen-s • github.com/SVSPraveen

SUMMARY

AI/ML Engineer specializing in production AI systems, Agentic AI, and Retrieval-Augmented Generation (RAG). Experienced in designing end-to-end LLM applications, multi-agent orchestration, hybrid search, workflow automation, and scalable backend services using Python, FastAPI, React, and modern AI infrastructure. Built and shipped production-oriented AI solutions spanning healthcare (RespiRAG) and autonomous job-search automation (SPrav™ Job AI) — including a publicly deployed, downloadable AI product with 28,700+ tracked job sources. B.Tech CSE, Amity University (2022–2026). Based in Mumbai, Maharashtra.

CORE SKILLS

AI / AGENTIC SYSTEMS LangGraph, LangChain, Agentic RAG, Multi-Agent Systems (Mixture-of-Experts routing), Model Context Protocol (MCP), Tool-Calling, LLMs (Llama 3.3, Qwen 2.5, DeepSeek-R1, GPT-OSS-120B), LoRA Fine-tuning, LlamaIndex, Self-RAG, CRAG

RETRIEVAL & SEARCH Hybrid Search (BM25 + Dense), Reciprocal Rank Fusion, Cross-Encoder Re-ranking, Cosine Similarity Scoring, Semantic Caching, Sentence-Transformers, HuggingFace Transformers

LLMS & VECTOR DATABASES Groq, Ollama, vLLM, OpenAI API, AWS Bedrock, Qdrant, pgvector, Pinecone, MedCPT, DeepSeek-R1/R2, Qwen 2.5/3, Llama 3.x

ML & EVALUATION Scikit-learn, LightGBM, XGBoost, TensorFlow/Keras, MLflow, RAGAS, RapidFuzz, NDCG, Hallucination Detection, Drift Detection, Citation Accuracy, pytest

FULL STACK, BACKEND & AUTOMATION FastAPI, REST APIs, React 18, JavaScript, TypeScript, Python, Playwright, PostgreSQL, MongoDB, SQLite, DuckDB, Redis, SQLAlchemy, Alembic, Pydantic, Docker, PyWebView

CLOUD, DEPLOYMENT & SECURITY AWS, Oracle OCI, GCP, Cloudflare Pages, Kubernetes, HIPAA, GDPR, FDA 21 CFR Part 11, OAuth2, JWT, AES-GCM, Prometheus, Linux, Git

EXPERIENCE

AI/ML Engineer Intern — MobcoderAI, Noida, UP (Onsite)

Apr 2026 – Jul 2026

- Built a hybrid BM25 + dense-vector retrieval pipeline with Reciprocal Rank Fusion and Cross-Encoder re-ranking on Qdrant Cloud for SEC 10-K/10-Q filing search across 4 companies, adding a semantic cache (cosine similarity ≥ 0.92) that served repeat queries with zero added LLM latency.
- Designed a multi-agent orchestration system — 6 specialized agents plus a router for intent decomposition, comparison, risk analysis, and trend queries — that dynamically routed work between Llama 3.1 8B and Llama 3.3 70B on Groq to balance cost, latency, and accuracy.
- Built a deterministic, LLM-free evaluation harness (RapidFuzz) scoring Faithfulness, Groundness, Hallucination Rate, and Citation Accuracy, gating numerical claims to a 1% tolerance against source filings, tracked end-to-end in MLflow.

TECHNICAL PROJECTS

SPrav™ Job AI — Autonomous Career Intelligence & Agentic Job-Application Platform

github.com/SVSPraveen/SPrav-Job-AI • sprav-job-ai.pages.dev

Python · FastAPI · React 18 · SQLite · Ollama · Groq · Playwright · PyWebView · Redis · ReportLab · Cloudflare Pages

- Designed and shipped a local-first, multi-agent pipeline (Discovery → Extraction → Fit Scoring → Tailoring → Verification → Apply) that routes tasks across specialized models (Qwen 2.5, DeepSeek-R1, GPT-OSS-120B via Groq with a local Qwen 2.5-Coder fallback, bespoke-minicheck, nomic-embed-text) and cuts hallucinated resume claims via an automated verifier feedback loop.
- Built a Playwright-based auto-apply engine that submits tailored applications directly to Greenhouse, Lever, Ashby, Workday, and SmartRecruiters, enforcing 3-tier circuit breakers (per-company/per-portal/global daily caps) to keep automation within safe rate limits.
- Engineered deterministic project selection, AI-generated content verification against a personal knowledge base, encrypted local credential storage (AES-GCM/XOR with a master recovery key), and hybrid Groq/Ollama inference tuned to run within an 8GB VRAM ceiling via sequential GPU thread-locking.
- Packaged and deployed the app as a portable Windows release (v2.4, ~385MB) and shipped a public landing page on Cloudflare Pages with an interactive in-browser ATS match simulator, product screenshots, and a walkthrough demo, scoring candidates against 28,700+ tracked job sources.

RespiRAG — Clinical Precision AI for Oncology (NSCLC)

github.com/SVSPraveen/RespiRAG • Live Demo: pharma-spine-ai.vercel.app

Python · FastAPI · React 18 · Qdrant · MedCPT · PostgreSQL · Redis · Groq · Docker · JWT

- Built a governed RAG pipeline over 600+ FDA drug labels and 3,900+ pages of NCCN clinical guidelines, combining HyDE query expansion, MedCPT biomedical embeddings, Cross-Encoder re-ranking, score gating (rejecting chunks below a 0.15 relevance threshold), and an LLM-as-judge verifier to reduce hallucinations.
- Implemented JWT-based RBAC across Patient/HCP/Admin roles with refresh-token rotation and Redis session revocation, guarded by a governance layer that blocks jailbreak and prompt-injection attempts.
- Reduced redundant vector-DB and LLM calls with a Redis citation cache and multi-key Groq rotation with Ollama fallback for graceful rate limiting.
- Built a CI/CD-compatible offline evaluation harness and a document version-control system that flags superseded clinical literature.

EDUCATION & CERTIFICATIONS

B.Tech, Computer Science & Engineering — Amity University, Uttar Pradesh • 2022 – 2026 • CGPA 7.4

Cloud Certifications: Oracle OCI Multicloud Architect Professional, OCI Networking Professional, OCI Foundation Associate, Google Cloud Computing Foundations

Security Certifications: Cisco Ethical Hacker, IBM Cybersecurity Analyst, Saviynt Identity Security for AI Age