Sarayu S.

AI / GenAI Engineer — RAG, LLM & Data Platforms

rayusar7@gmail.com | +91 75691 28363 | linkedin.com/in/sarayu-s-a43565292 | ssarayu-portfolio.netlify.app | Tirupati, India | Open to Remote / Hybrid

PROFILE

GenAI engineer with **3 years** designing, building and shipping production LLM systems for a Fortune-500 medical-device client (Medtronic, USA). Owns RAG architecture end to end — retrieval design, vector search, guardrails, evaluation and orchestration — including an NL-to-SQL engine that lifted query accuracy from **70% to 95%** across 30+ PostgreSQL schemas, and a regulatory RAG platform that auto-generates ISO 10993 biocompatibility evaluation reports. Backed by a strong data-engineering foundation (Prefect DAG orchestration, metadata mapping, transformation rules, automated validation) and the domain fluency to gather requirements directly from US regulatory stakeholders.

CORE SKILLS

GenAI & LLM	RAG Architecture · LangChain · Azure OpenAI / OpenAI / Anthropic APIs · Prompt Engineering · LLM Evaluation & Guardrails · Semantic Chunking · Hybrid Search · HuggingFace
Data & Storage	Python (Pandas, NumPy) · SQL · PostgreSQL · MongoDB · Redis · Milvus · HNSW Vector Search · Dimensional Modeling · ERD Design
Pipelines	Prefect Workflow Orchestration (DAG-based, transferable to Airflow) · ETL · Metadata Mapping · Transformation (TR) Rules · Data Validation & Quality Gates · Retry / Observability
Backend & Cloud	FastAPI · Flask · REST APIs · Docker · HPC Deployment · Azure · AWS · GCP · Git
Delivery & BI	Agile / Scrum · Polarion ALM (Siemens) · Zoho Sprints · UAT Support · Requirement Gathering · Tableau · Power BI · KPI Design · Data Visualization · ISO 10993

EXPERIENCE

Data & GenAI Analyst — Exafluence — Client: Medtronic, USA — Sep 2023 – Present · 3 yrs

- **ResearchAI — NL-to-SQL query engine.** Built and productionised a natural-language-to-SQL engine (Python, LLMs, Milvus, PostgreSQL) over **30+ regulatory schemas**, raising retrieval accuracy from **70% to 95%** through schema-aware retrieval, few-shot prompt tuning and an evaluation harness; cut analyst SQL turnaround from hours to seconds.
- **DocSeek — RAG retrieval platform.** Engineered an HNSW vector-search system with hybrid keyword matching, multi-attribute chunking and query rewriting over ISO 10993 regulatory reports, reducing document search time by **~70%** for US regulatory teams.
- **DocCreate — automated BER generation.** Architected a Biological Evaluation Report generation platform on Flask, MongoDB and Milvus with a three-mode strategy engine (Template, RAG-Generate, Manual-Required), retrieval-guardrail enforcement, draft-retrieval audit logging and native DOCX export — replacing a manual authoring workflow for regulatory writers.
- **Prefect workflow orchestration.** Designed and operated end-to-end DAG pipelines in Prefect covering ingestion, **metadata mapping**, **transformation-rule (TR) application**, automated **data validation** gates and downstream publishing — with task-level retries, dependency management and run observability; equivalent in pattern to Airflow DAG orchestration.
- **E&L compound classification.** Built a batch LLM pipeline (OpenAI API + PubChem) classifying **900+ extractables/leachables** against ISO 10993 toxicological thresholds with automated CAS-number validation, replacing manual literature lookup.
- **Deployment & platform.** Containerised and deployed a multi-extension Flask application on a Dockerised HPC cluster; debugged security-middleware and runtime failures to restore production availability.
- **Delivery & issue management.** Ran weekly requirement-gathering calls with US stakeholders; **raised, tracked and resolved defect and change-request tickets in Polarion (Siemens ALM)**, owning assigned items from investigation through fix to closure, and supported **UAT** cycles under Agile/Scrum.

SELECTED PROJECTS

BioComp / cwan Platform — Flask · Docker · HPC · Milvus · MongoDB. Multi-extension regulatory AI platform hosting DocSeek, DocCreate and ELMapper behind a shared retrieval and document layer; includes a Gap Analysis module scoring RAG-generated report sections against source evidence.

SkillClimb — Gamified Upskilling Platform — React · Vite · Firebase · Firestore · Playwright · Lighthouse. Two arenas (Code Climb, Mind Climb) across 12 tracks with level progression, badges, timed Interview Mode and a real-time leaderboard; validated with E2E tests and performance audits.

The Vault — MBA Revision Portal — HTML/CSS/JS · Vercel. Six-subject exam-prep portal (MCQ cheat sheets and descriptive answer guides) built and deployed for an entire MBA cohort.

EDUCATION

MBA, Ai & Data Science — SRM University — **In progress · Expected 2027**

MSc, Biotechnology — Sri Ramachandra Institutions (High Distinction) — **2021 – 2023**

BSc, Microbiology & Biochemistry — Mohan Babu University (High Distinction) — **2018 – 2021**

CERTIFICATIONS

Generative AI with LLMs — DeepLearning.AI & AWS (2025) · Infrastructure Modernization — Google Cloud (2025) · MongoDB SI Associate — MongoDB University (2025)