

Akshit Rampershad

“Turning enterprise operations into scalable AI-native systems”

813-397-9209 | akshitrampershad@gmail.com | [LinkedIn](#) | [GitHub](#) | San Francisco, CA | Open To Relocate

EXPERIENCE

AI Engineer | Cambrian Lab, Pleasanton, CA

May 2026 – Present

- Architected an end-to-end **Product Change Notification (PCN) to Manufacture Part Number (MPN) matching system** using a **4-leg hybrid retriever** (exact, fuzzy, prefix, vector) with an **LLM ranking pass** and a 3-band score+margin policy.
- Optimized vector search to a flat 0.36s match latency on a 70K+ corpus, **embedding 4 manufacturing fields with OpenAI text-embedding-3-small 1536-d into unified cosine Qdrant collection** with keyword payload indexes scoping HNSW topK.
- Achieved 0 false auto-confirmations across 100+ validation runs by gating HNSW top-K retrieval candidates on score and margin thresholds, **auto-confirming high-confidence matches and routing ambiguous ones to human steward review**.
- Eliminated manual assembly / finished good tracing by building a recursive where-used engine with parent-traversal graph algorithms and cycle-detection guards, **automating downstream Bill of Materials (BOM) impact analysis in SAP**.
- Orchestrated **29+ daily supplier change intakes through PcnAutoIntake**, an LLM pipeline chaining structured extraction, MPN matching and BOM impact analysis behind confidence-gated review, **cutting analysis from 2hr to 10sec per notice**.

Data Engineer | 22nd Century Technology Inc. McLean, VA

Sept 2025 – Mar 2026

- Led **technical architecture and delivery of a governed enterprise RAG system**, translating complex multi-million dollar federal/state proposal requirements into a scalable **knowledge-management platform** across 20K+ document corpus.
- Engineered a hybrid ingestion pipeline using Python, PySpark, and Graph API, indexing **distributed SharePoint content into a 1,024-dim Qdrant vector database** with **BGE-Large-EN-v1.5 embeddings** and section-aware document chunking.
- Designed LlamaIndex-based retrieval pipeline **narrowing 30 candidates to 8 via 5-field metadata filtering, vector search, and Cross-Encoder re-ranking**, feeding LangChain orchestration to return RFP-scoped drafts grounded in prior wins.
- Validated retrieval against a held-out benchmark of 20 proposal queries, achieved a 75%+ hit rate with a 54% first-pass draft acceptance rate, **reducing proposal draft assembly time from 7 days to under 4 hours across 20+ proposals**.
- Facilitated delivery across 4 cross-functional teams for 8 sprints, authored enterprise AI adoption runbooks, and led enablement sessions for proposal writers, **achieving 88% sustained weekly adoption across 25+ cross-functional users**.

Data Engineer | Surge Infotech, Boston, MA

Sept 2024 – Sept 2025

- Engineered **high-throughput geospatial pipelines feeding an ML site-suitability model** on Databricks AutoML+MLflow, scoring parcel buildability against zoning rules and geotechnical constraints at 72% accuracy vs site feasibility assessments
- Built **FastAPI data extraction services ingesting 50M+ parcel and zoning records into AWS S3**, leveraging Databricks Auto Loader for incremental file discovery and schema inference and DLT for governed ingestion and automated quality checks.
- **Designed and orchestrated a multi-stage Medallion data architecture (Bronze/Silver/Gold)** on Databricks LakeFlow, establishing automated schema evolution and real-time quality validation gates across **spatial and zoning records**.
- Developed **OpenAI GPT-4-based reporting engine that synthesized pipeline and model metrics into executive-ready business intelligence summaries**, replacing manual reporting and accelerating C-suite strategic planning cycles by 34%.
- Aligned foundational model capabilities with commercial demands, **presented smart city market adoption trends to executive leadership to secure multi-quarter budget buy-in for the core product roadmap**.

Business Data Analyst | UKG, India

Oct 2020 – Aug 2022

- Led business and technical requirements analysis for UKG TALK post-acquisition, **identifying key product gaps and shaping our strategic data backed product roadmaps**, successfully supported 10,000+ active enterprise users worldwide.
- Detected **on-boarding drop-offs points using SQL funnel analysis and Python cohort segmentation**, executing process optimization techniques that boosted premium **feature monthly conversions rate by 29%** and **trial to paid rate by 25%**.
- Designed **JIRA defect tracking dashboards monitoring production/staging issues**, enabling 4 cross-functional teams to unblock dependencies and critical bugs, **resulting a 43% improvement in phased sign-offs with 94% sprint delivery rate**.
- Developed **interactive Tableau/Power BI dashboards using advanced parameters and calculated fields** to report real time campaign ROI and segment conversion metrics to non-technical stakeholders during executive quarterly reviews.
- Led **cross-functional UAT loops, post-production checks, and product marketing launch documentation** to control critical integration friction for high-value Fortune 500 clients, **boosting platform adoption rates by 2X within 12 months**.

EDUCATION

- **MS in Business Analytics and Information Systems**, University of South Florida, Tampa, Florida **May 2024**
- **BS in Computer Science and Engineering**, Gandhi Institute of Technology and Management, India **July 2020**

SKILLS

- **Agentic AI & LLM Orchestration:** Anthropic Claude API, OpenAI GPT-4, Gemini-3.7-flash, LangChain, LlamaIndex, MCP, Prompt Engineering, Data Extraction, Tool Orchestration, Multi-System Agent Workflows, Autonomous Agentic Engines.
- **RAG & Retrieval:** Hybrid Retrieval (BM25/dense/RRF), Qdrant, Vector Search (HNSW/ANN), 2Stage Retrieval, Embedding (text-embedding3small/g-embedding001/BGELarge), Entity Resolution, Cross-Encoder Rank, Payload Indexing, Chunking.
- **AI Evaluation & Reliability:** Confidence Calibration, Threshold/Margin Policy, Validation Harnesses, False-Positive Control, Human-In-The-Loop, Latency, Agent Tracing & Observability (LangSmith), Token & Cost Tracking, Latency Benchmarking.
- **Data Engineering & Cloud:** Python, PySpark, SQL, Java, FastAPI, Databricks (Auto Loader, DLT, Lakeflow, Medallion Architecture), Snowflake (Snowpipe, Streams & Tasks, Snowsight), MySQL, AWS S3/EC2, MLflow, AutoML, Git/Bitbucket.
- **Solution Engineering:** Technical Feasibility Analysis, Product Gap Analysis, Enterprise Application Integration, Proofs Of Concept (POCs), Solution Delivery, Product/Project Management, Stakeholder Management, Technical Documentation.

ACCOMPLISHMENT

- **Certifications:** Model Context Protocol ([Anthropic](#)), Claude with Amazon Bedrock ([Anthropic](#)), Claude with Google Cloud's Vertex AI ([Anthropic](#)), Building with Claude API ([Anthropic](#)), Introduction to agent skills ([Anthropic](#)), Introduction to subagents ([Anthropic](#)), Snowflake Data Engineering ([Snowflake](#)), McKinsey Forward Learning ([Mckinsey.org](#)).
- **Recognition:** Exceptional Performer Award (UKG -2022), Technical Excellence Awards (UKG -2021).
- **Leadership:** Graduate Assistant (USF-2024), Data Science Club (USF-2023), PMI Volunteer (Tampa-2022).