Sai Darshini Amudala Suresh

Machine Learning Engineer — Content Understanding, LLM & Search Relevance

darshiniamudala09@gmail.com linkedin.com/in/darshiniamudala +1 (314) 335-0736

Positioning: ML engineer specializing in **LLM-powered content understanding**, **semantic retrieval (RAG)**, and **search relevance/metadata intelligence**—shipping production NLP systems with measurable impact and strong ownership.

Experience

Kaiser Permanente — AI/ML Engineer Feb 2024 – Present
Oakland, CA (Hybrid)

- Owned and shipped **LLM/NLP pipelines** that transform large-scale unstructured text into **searchable, structured knowledge** (summarization, entity/attribute extraction, semantic indexing) for downstream decision and discovery workflows.
- Built and productionized **retrieval-augmented generation (RAG)** systems for high-precision knowledge access, improving response relevance and reducing manual lookup overhead in real-time workflows.
- Developed transformer-based summarization and understanding models (BERT/BioBERT-class), improving documentation efficiency by **65%** while maintaining quality and latency constraints.
- Designed end-to-end ML pipelines (data → features → training → serving → monitoring), enabling high-scale inference and reliable continuous improvement loops.
- Delivered measurable outcomes by optimizing model quality trade-offs and deployment constraints, contributing to **42%** improvement in early detection signals and **28%** reduction in readmission-related outcomes (pipeline-driven impact).
- Partnered cross functionally with product/engineering stakeholders to integrate ML outputs into real workflows prioritizing **clarity, reliability, and user trust** (explainability + monitoring).

Accenture — AI/ML Engineer Jun 2021 – Jul 2023
Bangalore, India

- Built **text intelligence pipelines** to extract sentiment, entities, and semantic signals from news/market sources; operationalized signals for relevance-aware decision systems.
- Developed ML models to prioritize and rank information streams under real-time constraints, contributing to **18%** improvement in portfolio decision outcomes.
- Implemented distributed training/large-scale processing (Spark + scalable ML tooling), reducing iteration time from days to hours for complex model workloads.
- Automated feature engineering on **100TB+** datasets using PySpark/Hadoop ecosystems, improving training throughput and reliability for production pipelines.
- Deployed containerized ML microservices with high availability (**99.9% uptime**), with CI/CD and monitoring practices to support continuous delivery.

Sri Venkateswara College of Engineering — Research Assistant Jan 2021 – May 2021
Tirupati, India

- Co-authored IEEE ICCCI 2022 paper on resource allocation in D2D communications; improved spectral efficiency and reduced outage probability by **20%**.

Selected Projects (Netflix-Aligned)

- **RAG Knowledge Retrieval System** — Designed retrieval + generation workflow using embeddings and a vector index to improve high-precision question answering over internal documentation; implemented evaluation strategy for relevance and hallucination control.
- **LLM-Powered Summarization & Semantic Enrichment** — Built summarization + structured extraction pipeline to turn long-form notes into concise outputs and normalized entities for search and discovery.
- **Multi-Modal Signal Fusion (Text + Structured Data)** — Combined unstructured text signals with structured records for improved downstream prediction and relevance scoring; optimized for robustness and monitoring in production.

Core Skills

Content Understanding & LLMs: LLM systems, embeddings, semantic search, summarization, entity extraction, prompt iteration, evaluation strategies
Retrieval & Search: RAG pipelines, vector search, relevance optimization, offline/online quality thinking
ML Engineering: production ML pipelines, model monitoring, drift awareness, CI/CD patterns, scalable data processing
Tools: Python, SQL, Spark/PySpark, Hugging Face, LangChain, AWS (SageMaker/EMR/Lambda), GCP (Vertex AI/BigQuery), Docker, Kubernetes, MLflow

Education

Saint Louis University — M.S. Information Systems (CGPA: 3.6/4.0) St. Louis, MO
Sri Venkateswara College of Engineering — B.Tech Electronics & Communication (CGPA: 7.89/10) Tirupati, India

Certifications

AWS Certified Machine Learning – Specialty • Google Cloud Professional Data Engineer • Professional Scrum Master I (PSM I)