Saketh Velidimalla

623-254-8511 sakethv7@gmail.com https://www.linkedin.com/in/sakethvelidimalla/ https://github.com/Sakethv7 Portfolio

SUMMARY

Data Scientist / **Machine Learning Engineer** with experience deploying LLM-driven decision systems at enterprise scale. Strong background in Python, distributed data pipelines, model evaluation, and post-production reliability for real-world AI systems.

PROFESSIONAL EXPERIENCE

Data Scientist, Johnson & Johnson | New Brunswick, New Jersey | **February 2025 - Present**

- Built and deployed production RAG based LLM applications serving **~140K employees**, handling **10K–25K** monthly queries and deflecting 30–40% of Tier-1 HR policy tickets through automated self-service.
- Designed and operated **RAG pipelines** over **10K+ HR policy documents**, using Qdrant vector DB for retrieval, reducing hallucinated or outdated responses by ~45% versus prompt-only baselines.
- Implemented **MCP-based agent routing** for a procurement GenAI workflow, dynamically deciding between RAG-based information retrieval and agentic actions (e.g., update, remove, or view Purchase Orders) by clarifying user intent and orchestrating downstream system calls.
- Established content governance and security guardrails, including trusted-source filtering and red-teaming for prompt injection and policy bypass, reducing stale or non-compliant responses by ~60%.
- Performed **parameter-efficient fine-tuning** (LoRA / instruction tuning**)** on LLaMA-3 models, improving policy adherence and response quality by ~18–22%.
- Integrated Arize Phoenix for end-to-end LLM observability, tracing RAG workflows and monitoring latency and quality patterns across production traffic.
- Built Power BI dashboards for leadership tracking **60K+ user queries**, adoption trends, and recurring policy topics, helping reduce repeated knowledge requests and organizational brain drain.
- Led **post-production incident triage and RCA**, resolving business-level risk and user issues and reducing repeat GenAI incidents by **~35%**

Data Analytics Engineer, iDwTeam LLC | Alpharetta, Georgia | **November 2024 – February 2025**

- Automated SAP deployment workflows using GitLab CI/CD pipelines, eliminating **~27% of redundant manual steps and approval screens**.
- Reduced average SAP deployment cycle time from **~3.5 hours to ~2.5 hours** across five enterprise environments.
- Built Tableau dashboards tracking deployment success rates, failures, and lead times to improve release visibility for operations teams.

Data Analytics Engineer, Hewlett Packard Inc. | Spring, TX | **July 2024 – November 2024**

- Optimized PySpark pipelines on AWS EMR, reducing end-to-end processing latency by **~25%** across large-scale system telemetry datasets.
- Improved pipeline reliability by **~30%** through Airflow orchestration, retries, SLA monitoring, and alerting.
- Built Power BI dashboards on Redshift to surface system health metrics and performance trends used in recurring engineering reviews.
- Analysed hardware performance signals (power usage, BIOS utilization, time-of-flight) to identify underperforming configurations and optimization opportunities.
- Collaborated with Data Engineering teams on **code reviews and production debugging**, reducing repeat analytics defects and rework.

Machine Learning Engineer, ECrent Worldwide Company | Bengaluru, India | **July 2021 – June 2022**

- Built end-to-end ML systems for real-estate pricing and personalized recommendations, influencing **~15–20% of booking flows**.
- Improved recommendation relevance by ~12–18% (precision@k) using hybrid collaborative and content-based models.
- Reduced pricing prediction error (MAPE) by **~10%** through feature engineering and gradient-based optimization.
- Developed NLP pipelines using Word2Vec and sentence embeddings, improving semantic matching by **~20%**.
- Deployed models via Flask REST APIs supporting real-time inference with sub-second latency.
- Validated improvements through A/B testing, identifying statistically significant gains in engagement and conversion metrics.

EDUCATION

Arizona State University **Arizona, USA**
Master of Science in Information Technology (Data Science) **GPA 4.0/4.0**

TECHNICAL SKILLS

- **GenAI & LLM Systems:** RAG Pipelines • Agentic Workflows (MCP) • Prompt Engineering • OpenAI API (text-embedding-3-large) • Gemini API • Claude API • Arize Phoenix • Red Teaming & Guardrails • RASA NLU
- **Machine Learning & AI:** PyTorch • TensorFlow • Hugging Face Transformers • Scikit-learn • AWS SageMaker
- **Programming & APIs:** Python • SQL • RESTful APIs • Flask • JSON • YAML
- **Data Engineering & Pipelines:** PySpark • Apache Airflow • ETL Pipelines • Data Modeling • Hadoop • Hive
- **Cloud & DevOps**: AWS (S3, EC2, EMR, Redshift, EKS, Lambda, DynamoDB, CloudWatch) • Docker • Kubernetes • Azure DevOps
- **Databases & Vector Stores:** Qdrant • PostgreSQL • MongoDB • Snowflake • Redshift • DynamoDB
- **Analytics & BI:** Power BI (DAX) • Tableau • Observability (Arize-Phoenix)
- **Tools & Practices:** Git • Jira • Confluence • Agile/Scrum • VS Code • IntelliJ