Shashi Kiran C

Fairfield, CT, USA | c.shashikiran421@gmail.com | +1 (203) 243 – 0469
https://github.com/kiransview | https://www.linkedin.com/in/shashikiranc/ | https://www.kiransperspective.com/

SUMMARY

Senior AI Engineer and Generative AI Specialist with 8+ years of experience building enterprise-scale AI/ML solutions across domains such as document intelligence, RAG architecture, and predictive analytics. Expert in LLM integration, model fine-tuning (LoRA/QLoRA), hybrid retrieval pipelines, and real-time inference systems. Proven ability to bridge business needs with AI innovation leveraging semantic search, orchestration frameworks (LangChain, LangGraph), and cloud-native tools (AWS, GCP) to deliver scalable, context-aware solutions. Actively seeking AI-focused roles at the intersection of Generative AI, retrieval systems, and enterprise intelligence.

EXPERIENCE

Lantern Point Labs, Fairfield, Connecticut, USA
Data & AI Lead **(2025 – Present)**

- Designed and deployed a Knowledge Management System using **ChromaDB + Langchain** locally, helping founders understand the capabilities of GenAI and RAG systems in a cost-effective proof-of-concept setting.
- Implemented an **Amazon Q** interface for **enterprise search** & integrated with Amazon S3 for scalable data storage and retrieval, extending with **Lambda** functions for efficient conditional document retrieval at scale resulting in 10% boost in retriever accuracy.
- Led a **team of 2** along with cross-functional teams to build a hybrid retrieval platform unifying enterprise documentation with personal productivity tools (Gmail, Slack, Calendar) into a personalized knowledge assistant using Amazon Q, boosting information access.
- Developed an **agentic research workflow for finance** with **LangGraph + LangChain** orchestrated over **Together AI serverless models**, conducting **automated financial data retrieval and synthesis**, reducing manual research time for consultants by over 40%.
- Conducted evaluation and optimization (using **Arize Phoenix**) of research agent, applying metrics (latency, accuracy, coverage) with **p95 of 15s and p50 of 2s**, improving agent response quality by 25% and directly supporting enterprise adoption.
- Planning and designed deployment of the **Research Agent as an Amazon Q App/Plugin** for seamless consultant access, with backend hosting on **AWS Lambda or EC2** for scalable orchestration.

Fairfield University, Fairfield, Connecticut, USA
Computer Vision and Machine Learning Research Assistant **(2024 – 2025)**

- Led Machine Learning, Generative AI (Agent), and Computer Vision research projects, improving model performance with advanced algorithms (Hugging Face, Llama, TensorFlow, PyTorch) for increased efficiency and accuracy compared to SOTA models.
- Built and optimized transformer-based NLP models (Mistral) for summarization and intent detection using Hugging Face, increasing text classification accuracy by 8% and reducing false positives by 12% across evaluation cyber security dataset.

Aligned Automation Services Limited India – Client: Dell (United States)
Senior Data Scientist & GenAI Engineer **(2022 – 2024)**

- Developed a feasibility analysis metric for Generative AI, Machine Learning, and automation solutions, improving implementation viability assessment accuracy by 8% and enhancing business impact evaluation.
- Developed and optimized generative AI models (Llama, GPT, Mixtral) inference with RAG using vector databases (Chroma DB, FAISS) and finetuning with LoRA and QLoRA, enhancing model performance and cutting costs by transitioning to open source.
- Delivered warehouse demand and supply forecasts by building time series models (Prophet, ARIMA, Croston), improving demand planning efficiency (accuracy) by 15% through advanced predictive analytics.
- Partnered with **Dell's strategy and operations teams** to translate business requirements into AI/ML solutions, producing measurable ROI through improved forecasting, automation insights, and generative AI adoption roadmaps.

Tata Consultancy Services India – Client: Bed Bath and Beyond (United States)
Data Scientist **(2021 – 2022)**

- GCP Workbench notebook interactions (jupyer lab) for quick experimentation and implemented Kubeflow to orchestrate Kubernetes clusters for pipeline architecture on Vertex Pipelines.
- Performed sales transference analysis to understand customer movement and purchasing shifts across nearby stores, improving forecast accuracy by 13% and product layout efficiency by 10% through market basket analysis and behavioral insights.
- Automated real-time reporting workflows using GCP and DOMO, reducing manual reporting time by 25% and enabling faster, data-driven decision-making across key business units and departments.
- Implemented market basket analysis and a rule-based recommendation engine to uncover product group patterns, leading to a 10% improvement in product layout efficiency across high-traffic retail zones (online and offline).

HCL Technologies Bengaluru, Karnataka, India – Client: Applied Materials (United States)
Senior Data Analyst **(2018 – 2020)**

- Increased sentiment analysis accuracy by 8% leveraging TensorFlow, Keras, and LSTM to classify customer sentiments based on field service reports, using Regex, NLTK, TFIDF, Word2Vec, and Gensim for NLP preprocessing and embedding.

Bosch Limited India
Data Analyst **(2016 – 2018)**

- Developed and delivered high-priority dashboards using PowerBI and Tableau, achieving 100% on time delivery for critical feature implementations while maintaining logistics and buyer activity dashboards.

EDUCATION

Fairfield University, Fairfield, CT
Master of Science in Business Analytics (MSBA) **(2024 – 2025)**
Visvesvaraya Technological University, India – B.E in Mechanical **(2012 – 2016)**

RESEARCH PUBLICATION

EXPLORING THE DEPTH OF THE KAN METHOD FOR HYPERSPECTRAL IMAGE CLASSIFICATION: KANs boost HSI classification by efficiently modelling spectral complexities. **ASEE-NE-25 BEST PAPER (First Position)**
LEVERAGING LARGE LANGUAGE MODELS FOR AUTOMATED DETECTION OF COOKIE AND SESSION MANAGEMENT VULNERABILITIES: Can LLMs enhance web security by detecting cookies and session vulnerabilities. **ASEE-NE-25 BEST PAPER (Third Position)**

ANALYTICAL & TECHNICAL SKILLS

Platforms: OpenAI, Langchain, Langgraph, GitHub Copilot, RAG (Vector Stores – FIASS, Chroma dB, Pinecone), Hugging Face, Mistral.
Models: Regression, Classification, Forecasting, Llama3, GPT 3.5 turbo, Mixtral 8x7b, Open Source (Hugging Face), Deep Learning Architectures (Transformers, Encoder Decoder Architecture, ImageNet)
Tools/Software: Python (Pandas, Numpy), RStudio, SQL, Tableau, PowerBI, Microsoft Excel, NLP, Databricks, PyTorch, TensorFlow, Computer Vision, databricks
Storage: Teradata, Google Cloud BigQuery, Data Lake
Web: Microsoft Azure, Google Cloud Platform, AWS Sage Maker