Contact

+1 334-294-9548
koukuntlarakeshr@gmail.com
Portfolio

RAKESH REDDY

GEN AI ENGINEER

Professional Summary

- Professional with 10+ years of experience, specializing in Generative AI (Gen AI), Large Language Models (LLMs), and MLOps across cloud (AWS, Azure, Oracle AI) and on-prem environments.
- Designed and deployed Gen AI infrastructure, optimizing GPU clusters (H100, A100, V100) for LLM training and inference.
- Developed scalable LLM solutions, fine-tuning models like GPT-4, Claude, LLaMA, Mistral, Falcon, and T5 using LoRA, QLoRA, and PEFT for cost-effective adaptation.
- Engineered retrieval-augmented generation (RAG) pipelines with FAISS, Pinecone, ChromaDB, Milvus for enhanced AI-powered search and knowledge retrieval.
- Built multimodal AI models integrating text, image, speech, and video using BLIP, CLIP, Whisper, ControlNet, and Stable Diffusion.
- Designed and implemented end-to-end AI/ML ecosystems, integrating LLMs, MLOps, cloud infrastructure, and scalable data pipelines to drive enterprise AI innovation.
- Implemented distributed training (DeepSpeed, ZeRO, FSDP) and optimized LLM inference with quantization (GPTQ, AWQ, FP8, INT8) for performance and cost-efficiency.
- Developed AI-powered chatbots & virtual assistants using LangChain, RAG, and knowledge graphs, enabling intelligent enterprise automation.
- Designed and automated LLMOps pipelines for model training, evaluation, deployment, and monitoring across AWS, Azure, and Oracle AI ecosystems.
- Integrated LLM-driven applications with FastAPI, Flask, gRPC, and Triton Server, deploying insights via Power BI, Oracle Analytics, and Tableau.
- Ensured AI security, governance, and compliance (HIPAA, GDPR, SOC 2, ISO 27001) through Explainable AI (XAI), SHAP, LIME, and privacy-preserving AI techniques.
- Developed AI-enhanced search and recommendation systems, leveraging vector embeddings and semantic search for enterprise applications.
- Led cross-functional teams, collaborating with data scientists, DevOps engineers, and product managers to deliver AI-driven business solutions.
- Researched and implemented cutting-edge Gen AI advancements, contributing to open-source AI projects and optimizing LLM scalability and efficiency.

Education

Bachelor of Technology
- Annamacharya Institue of Technology and Science

Skills

Programming: Python, Java, Go, Rust, SQL, NoSQL, JavaScript (React/Next.js), TypeScript

AI/ML: LLMs (GPT-4, Claude, LLaMA, Mistral, T5, BERT), Fine-Tuning (LoRA, QLoRA, PEFT), XAI (SHAP, LIME), Distributed Training(DeepSpeed, FSDP, ZeRO)

GenAI & LLMs: Prompt Engineering, RAG (FAISS, Pinecone, Milvus, Weaviate), AI Agents (LangChain, CrewAI, AutoGPT), Multimodal AI (CLIP,BLIP, Whisper, ControlNet)

MLOps & Deployment: LLMOps, CI/CD (MLflow, Kubeflow), Model Serving (vLLM, ONNX, TensorRT), AI Compliance (HIPAA, GDPR, SOC 2)

Cloud & Infrastructure: AWS, Azure, Oracle AI

Data & Databases: SQL/NoSQL, ETL (Airflow, Spark, Snowflake), Pandas, NumPy

APIs & Microservices: REST, FastAPI, Flask, gRPC, API Gateway, Serverless (AWS Lambda)

DevOps & Automation: Docker, Kubernetes, Terraform, ELK Stack, Monitoring (Prometheus, Grafana)

Visualization & Analytics: Power BI, Oracle Analytics, Tableau

Work Experience

Client: K Health, NY **January 2024 – To Date**
Role: Gen AI Engineer

Responsibilities:

- Designed and deployed Gen AI infrastructure on on-prem & cloud (AWS, Azure, Oracle AI), optimizing GPU clusters (H100, A100, V100) for LLM training & inference.
- Configured LLM-serving frameworks (vLLM, DeepSpeed, TensorRT, ONNX, Hugging Face Inference) for scalable, low-latency deployments.
- Implemented distributed training using FSDP, ZeRO, DeepSpeed, optimizing large-scale model fine-tuning.
- Developed vectorized search solutions (FAISS, Pinecone, ChromaDB, Weaviate, Milvus) for high-performance Retrieval-Augmented Generation (RAG).
- Optimized tokenization & embeddings to improve LLM-powered search retrieval & contextual understanding.
- Fine-tuned LLMs (Llama, Falcon, GPT, Mistral, Claude, T5, BERT) using LoRA, QLoRA, PEFT for cost-efficient domain adaptation.
- Built multimodal AI models integrating text, image, speech & video data using BLIP, CLIP, Whisper, ControlNet, Stable Diffusion.
- Developed AI-powered chatbots & virtual assistants using LangChain, RAG, and knowledge graphs for enhanced enterprise AI interactions.
- Engineered prompt tuning & instruction fine-tuning strategies for LLM accuracy, coherence, and bias reduction.
- Designed & automated LLMOps pipelines for model training, evaluation, deployment & monitoring across AWS, Azure, and Oracle AI.
- Optimized LLM inference with quantization techniques (GPTQ, AWQ, FP8, INT4, INT8) to improve latency & efficiency.
- Developed AI agents using LangChain, CrewAI, and AutoGPT, integrating multi-agent workflows into enterprise solutions.
- Deployed & integrated LLM-based services via FastAPI, Flask, gRPC, Triton Server, and integrated with Power BI, Oracle Analytics, and Tableau.
- Implemented AI security & compliance with Explainable AI (XAI), SHAP, LIME, ensuring HIPAA, GDPR, SOC 2, and ISO 27001 adherence.
- Applied privacy-preserving AI techniques (Federated Learning, Differential Privacy, Secure Aggregation) for enterprise & healthcare AI.
- Researched & implemented the latest Gen AI & LLM advancements (GPT-4, Claude, Gemini, Mixtral, Mistral, Llama, Falcon, T5).
- Contributed to open-source AI projects, optimizing Gen AI frameworks & multimodal NLP models for cutting-edge AI applications.

Client: TD Ameritrade, NE **July 2020 – December 2023**
Role: AI/ML Engineer

Responsibilities:

- Designed end-to-end AI/ML workflows, from data ingestion to model deployment, ensuring scalable and efficient solutions.
- Fine-tuned and optimized models like GPT-4, Claude, LLaMA, and Mistral using LoRA, QLoRA, and PEFT for domain-specific applications.
- Applied prompt tuning strategies to improve LLM accuracy, mitigate hallucinations, and optimize AI-generated outputs.
- Built AI-driven multi-agent systems using LangChain, CrewAI, and AutoGPT for task automation and intelligent decision-making.
- Designed scalable RAG pipelines with FAISS, Pinecone, and Milvus to enhance knowledge retrieval in AI-powered applications.
- Developed cloud-native MLOps pipelines (AWS SageMaker, Vertex AI, Azure ML) for seamless model training, deployment, and monitoring.

- Optimized large-scale AI models with DeepSpeed, FSDP, and ZeRO to maximize GPU efficiency and minimize costs.
- Integrated OpenAI's GPT models to enhance LLM-driven applications, leveraging APIs for chatbots, content generation, and enterprise automation.
- Optimized and deployed AI models on Azure ML, leveraging Azure Machine Learning pipelines for scalable training, deployment, and monitoring.
- Engineered semantic search solutions and recommender systems using vector embeddings for enhanced information retrieval.
- Integrated Azure Cognitive Services for NLP, speech recognition, and computer vision, enhancing AI-driven enterprise applications.
- Integrated XAI techniques (SHAP, LIME) and ensured AI governance, addressing fairness, bias, and regulatory compliance (HIPAA, GDPR, SOC 2).
- Developed AI-driven UIs with React/Next.js and deployed AI models as microservices using FastAPI, Flask, and gRPC.
- Deployed LLM inference with quantization (GPTQ, AWQ, INT8), reducing latency and improving resource efficiency.
- Developed models combining text, speech, and vision using CLIP, BLIP, Whisper, and Stable Diffusion.
- Worked with data scientists, DevOps engineers, and product teams while contributing to AI open-source initiatives.

Client: BlueOptima, India **May 2018 – October 2019**
Role: Data Scientist

Responsibilities:
- Experience with working on clickstream activities, Customer Journey activities, Fraud Detection, Sales and managing Store items.
- Used pandas, numpy, matplotlib, sci-kit-learn in Python for developing various machine learning algorithms.
- Experience with NoSQL databases such as MongoDB, Cassandra and Utilized SQL, NoSQL databases, Python programing and API interaction.
- Experience using ETL and data visualization tools like PowerBI.
- Implemented Classification using supervised algorithms like Logistic Regression, Decision trees.
- Data transformation from various resources, data organization, features extraction from raw and stored.
- Involved in defining the source to target data mappings, business rules, and data definitions.
- Performed automation engineer tasks and implemented the ELK stack (Elasticsearch, Kibana) for AWS EC2 hosts.
- Extracting the source data from Oracle tables, MS SQL Server, sequential files and other databases.

Client: BillDesk, India **February 2015 – April 2018**
Role: Software Developer

Responsibilities:
- Build and maintain tools and applications using Python, with integration to AWS services such as AWS Lambda, EC2, S3, and DynamoDB.
- Work with GitHub to integrate tools and applications, ensuring smooth development workflows.
- Write clean, efficient, and easy-to-understand code that is maintainable in the long term.
- Testing and Debugging: Troubleshoot, debug, and fix issues in the applications. Ensure the code works well by writing and running tests.
- Work on deploying applications and services to the AWS cloud using services like AWS Elastic Beanstalk, AWS S3, and AWS RDS.
- Help design and develop RESTful services and APIs for different applications.
- Work closely with other team members to create high-quality software solutions and meet project deadlines.
- Help design and develop RESTful APIs and microservices using AWS API Gateway.
- Contribute to enhancing tools, processes, and overall software quality.

Client: Perfios, India **June 2013 – January 2015**
Role: Software Developer

Responsibilities:

- Create HTML JSP forms for user input, & TreeGrid output
- Validate form input using Validator classes
- Used DAO layer concept for data transfer to DB
- Worked on Webservices using (SOAP/RESTful)
- Created database scripts to perform database updates
- Worked with XML files for mapping
- Worked with Wildfly & Jboss severs
- Implementing Servlets and JSP for designing purpose - using HTML & JavaScript.
- Developed the application front end: HTML forms & Java Page base Bean classes and Java Server Pages
- Used Log4j to create log files for troubleshooting