Praneeth Reddy Kesarapu

Generative AI Engineer

reddypraneeth066@gmail.com| +1-616-788-2345

LinkedIn: @praneethkesarapu1906| **GitHub**: @praneeethreddy | **Portfolio** : https://praneethreddy.netlify.app

PROFESSIONAL SUMMARY

AI/ML and Generative AI Engineer with **6+ years of experience** designing, building, and operating **production-grade machine learning, information retrieval, and LLM-powered systems** at enterprise scale. Proven expertise in **end-to-end RAG architectures, search and ranking systems, prompt engineering, and multi-agent workflows**, supporting millions of queries and thousands of users in real-world environments. Strong hands-on background in **NLP, deep learning, computer vision, and semantic search**, using **TensorFlow, PyTorch, Hugging Face, OpenSearch, and Elasticsearch**. Deep experience in **MLOps and platform engineering**, delivering scalable, observable, and secure AI services across **AWS, Azure, and GCP** using **CI/CD, Docker, Kubernetes, and cloud-native infrastructure**. Known for translating complex business and regulatory requirements into reliable AI systems with measurable performance, cost, and productivity impact.

SKILLS

Languages	Python, SQL, Java, Bash, Shell Scripting, R, C#
ML & Deep Learning	PyTorch, TensorFlow, Keras, Scikit-learn, XGBoost, LightGBM, CatBoost, NumPy, Pandas, SciPy, Statsmodels, NLTK, Hugging Face Transformers, Spacy, Matplotlib, Seaborn, Plotly, Gradio, Streamlit
Generative AI & LLMs	OpenAI (GPT-3, GPT-3.5, GPT-4, Codex, DALL·E, Whisper, CLIP, ChatGPT), Azure OpenAI, Anthropic Claude (v1–v3), Google Gemini, Meta LLaMA, Mistral, DeepSeek
LLM Frameworks & Tooling	LangChain, LangGraph, LlamaIndex, AutoGen, LangSmith, Transformers, Diffusers, PEFT, Accelerate
Fine-Tuning Techniques	LoRA, QLoRA, Parameter-Efficient Fine-Tuning (PEFT)
Prompt Engineering	Zero-shot, Few-shot, Chain-of-Thought, Function Calling, Custom Prompt Templates
RAG & Vector Databases	Pinecone, FAISS, Weaviate, Chroma, OpenSearch, Adaptive RAG, HyDE, CRAG
MLOps & DevOps	MLflow, Kubeflow, Argo Workflows, Airflow, Docker, Kubernetes, Jenkins, Azure DevOps, DVC
Monitoring & Observability	Prometheus, Grafana, TensorBoard, Azure Monitor, Application Insights
Cloud Platforms	**Azure:** OpenAI, Cognitive Search, Data Lake, Data Factory, App Service, Container Registry, Key Vault, Active Directory **AWS:** SageMaker, Bedrock, OpenSearch, S3, DynamoDB, Secrets Manager **GCP:** Vertex AI
APIs & Backend	FastAPI, Flask, Django, Django REST Framework
Conversational AI	Kore.ai, Azure Bot Services, IVR Integrations
Version Control	Git, GitHub, GitLab

Experience

Toyota, Plano, Texas **January 2024 – Present**

Role : Gen AI Engineer

Responsibilities:

- **Designed, deployed, and operated production-grade Generative AI platforms** serving **5,00,000+ enterprise users**, delivering real-time Q&A, document summarization, and decision-support across engineering and operations domains using Azure **Open AI**(GPT), **Gemini** and **Claude** Models.
- **Architected end-to-end Retrieval-Augmented Generation (RAG) pipelines** at enterprise scale over **10M+ documents**, implementing ingestion, chunking, embeddings, retrieval, and response synthesis using **Python, LlamaIndex, text-embedding models Amazon OpenSearch (Vector Engine), DynamoDB, and Amazon Bedrock.**

- **Built resilient ingestion frameworks** for unstructured and semi-structured data (PDFs, SharePoint, Excel, multimedia) using **Graph API, Tika, Tesseract OCR, ffmpeg, and custom metadata enrichment**, reducing content onboarding time by **50%+**.
- Integrated LLMs into **enterprise chat and voice platforms** including **Amazon Lex, Microsoft Teams, and IVR systems**, enhancing customer interactions with AI-driven insights.
- Fine-tuned **open-source LLMs (LLaMA, Mistral)** with **LoRA on Amazon SageMaker**, optimizing inference speed and accuracy using domain-specific datasets.
- **Designed and deployed multi-agent GenAI workflows** (retriever, validator, summarizer, reasoning, and citation agents) using **LangGraph**, improving answer grounding, traceability, and response accuracy in production environments.
- Built **scalable backend APIs** using **Python, FastAPI, and REST**, deployed on **Amazon Kubernetes Service (AKS)** for high-availability inference.
- Deployed GenAI workloads using **Amazon SageMaker, Amazon Bedrock, AWS Lambda, and Amazon ECS**, ensuring scalable, fault-tolerant, and low-latency inference pipelines.
- Implemented **CI/CD pipelines** using **AWS CodePipeline, CodeBuild, GitHub Actions**, and **Amazon ECR**, enabling automated, secure, and reproducible deployments.
- Enforced **enterprise data security and compliance** using **IAM roles and policies, AWS KMS, AWS Secrets Manager, VPC isolation, Private Link, Security Groups, and AWS Organizations**, maintaining governance and regulatory standards.
- Leveraged **PyTorch, TensorFlow, and Keras on Amazon SageMaker** for NLP models, predictive analytics, and fine-tuned LLM components.
- Evaluated **LLM and RAG workflow performance** using **RAGAS, DeepEval** assessing retrieval accuracy, response relevance, bias, and context consistency.
- **Implemented Model Context Protocol (MCP) based orchestration** to standardize context exchange between tools, retrievers, agents, and LLMs, enabling structured tool calling, deterministic agent behavior, improved observability, and consistent context management across multi-agent GenAI workflows.

Environment:

Amazon Bedrock, Amazon SageMaker, Amazon OpenSearch Service (Vector Search), Amazon EKS, Amazon ECS / AWS App Runner, AWS CodePipeline, AWS CodeBuild, AWS CodeCommit, Python, FastAPI, LangChain, LangGraph, LlamaIndex, PyTorch, TensorFlow, GitHub Actions, Docker

Omnicell, USA **August 2022 – December 2023**

Role: AI/ML Engineer

Responsibilities:

- Architected and deployed **end-to-end ML/DL pipelines** on **AWS** serving **100K+ predictions per month** with **<200ms P99 latency**, supporting risk analysis, document processing, and automation via **fault-tolerant REST APIs**, **auto-scaling infrastructure (EC2, Auto Scaling Groups, EKS)**, and **production monitoring (Amazon CloudWatch)**.
- Developed **computer vision solutions** using **CNNs and OpenCV**, processing **100K+ images monthly** stored in **Amazon S3**; implemented **OCR-based extraction** achieving **>92% accuracy**, reducing manual review efforts by **50%** across diverse document formats.
- Trained and optimized **deep learning models** using **TensorFlow, Keras, and PyTorch** on **GPU-enabled EC2 instances**, applying **transfer learning and custom architectures** to improve accuracy by **15–20%** while reducing inference costs by **~30%** through **quantization and pruning**.
- Implemented **MLOps best practices** using **Amazon SageMaker**, including **CI/CD pipelines (AWS CodePipeline)**, automated training/validation/deployment workflows, **model versioning**, **experiment tracking (MLflow)**, and **canary deployments**, reducing model release cycles from **weeks to days**.
- Built **real-time model monitoring and observability systems** using **SageMaker Model Monitor**, **CloudWatch**, **Prometheus**, and **Grafana** to track **data drift, prediction drift, performance degradation, data quality issues, and business KPIs**, enabling automated alerting and retraining.
- Designed robust **feature engineering and data preprocessing pipelines** using **Apache Spark on Amazon EMR / Databricks on AWS**, incorporating domain-specific transformations, data augmentation, and validation checks, improving model generalization by **~25%** and expanding effective training data by **300%**.

- Developed **high-throughput RESTful inference APIs** using **FastAPI and Flask**, deployed on **AWS Lambda with API Gateway** and **Amazon EKS**, implementing **batching, caching (Amazon ElastiCache/Redis)**, **asynchronous processing (Celery, RabbitMQ)**, and comprehensive request validation to serve predictions at scale.
- Implemented **distributed training** across **multi-GPU EC2 environments** using **PyTorch Distributed and TensorFlow MirroredStrategy**, along with **active learning frameworks**, reducing training time by **~60%** and labeling costs by **~45%** while maintaining model performance.
- Optimized **end-to-end inference pipelines** using **GPU acceleration**, **ONNX Runtime**, **batch processing**, and **auto-scaling**, reducing latency from **~800ms to <200ms** and improving automation throughput by **35–40%**.
- Built **containerized ML platforms** using **Docker and Amazon EKS**, integrating **feature stores (Feast)**, **data versioning**, **lineage tracking**, and **reproducible environments** across training, staging, and production.
- Collaborated with **cross-functional stakeholders** to translate regulatory and business requirements into **AWS-based ML solutions**, delivering **executive dashboards (Amazon QuickSight)** demonstrating **$500K+ annual cost savings** and enabling continuous feedback loops.

Environment:

AWS SageMaker, Amazon EC2 (GPU), Amazon EKS, AWS Lambda, API Gateway, Amazon S3, Amazon EMR, Databricks on AWS, AWS CodePipeline,Docker, Kubernetes, Apache Spark, FastAPI, Flask, TensorFlow, PyTorch, Keras, OpenCV, MLflow, ONNX Runtime, Prometheus, Grafana, Celery, RabbitMQ, Feast, SHAP, LIME, Amazon QuickSight

Accenture, India **January 2020 – August 2022**
Role: Data Scientist
Responsibilities:

- Built and deployed machine learning models for **classification, regression, and clustering** using **Scikit-learn, XGBoost, and TensorFlow** on scalable cloud infrastructure.
- Performed **exploratory data analysis (EDA)** and **feature engineering** using **Pandas, NumPy, and Matplotlib** to identify patterns and improve model performance.
- Designed and implemented **batch and real-time ETL pipelines** leveraging **Apache Spark, Hadoop, and Databricks** for large-scale data processing.
- Developed and productionized **NLP solutions** for sentiment analysis, text classification, and named entity recognition using **spaCy, NLTK, and Hugging Face Transformers**.
- Created and maintained **interactive dashboards and data visualizations** using **Power BI and Plotly** to deliver actionable insights to business stakeholders.
- Deployed machine learning models to production using **Flask and FastAPI**, integrating with enterprise systems through **scalable REST APIs**.
- Utilized **SQL** to query large datasets, perform complex joins and aggregations, and prepare optimized data views for ML workflows.
- Collaborated with **data engineers, analysts, and cross-functional teams** to deliver data-driven solutions for **healthcare and insurance** use cases.
- Applied robust **model evaluation and validation techniques** including confusion matrix, ROC-AUC, precision, and recall; conducted **hyperparameter tuning** to optimize performance.
- Documented end-to-end **data science workflows**, ensuring reproducibility, compliance, and adherence to data governance and security standards.

Environment:

AWS SageMaker, AWS Lambda, AWS Glue, AWS CodePipeline, AWS Kinesis Data Streams, Amazon EKS, Amazon CloudWatch, Docker, Prometheus, Grafana, FastAPI, TensorFlow, PyTorch, Hugging Face Transformers, MLflow, Optuna.

EDUCATION

Master's from University of Central Missouri, United States