SUNEEL KALVA

AI/ML Engineer | LLM Systems & Agents | RAG | Inference Optimization | MLOps

Dallas, TX (open to relocation) | (940) 233-6043 | suneel.kalva190700@gmail.com | linkedin.com/in/suneel2000 | github.com/suneel190700

PROFESSIONAL SUMMARY

AI/ML Engineer with 5+ years shipping production ML and generative AI systems end to end, from data pipelines and model training to low-latency serving, evaluation, and monitoring. Builds agentic AI systems including MCP servers, tool-calling orchestration, model routing, and agent simulation testing, with hands-on depth in RAG, supervised fine-tuning (LoRA/PEFT), and LLM inference optimization. Track record of measurable impact at scale: retrieval assistants over 500K+ documents, models trained on 10M+ records, and real-time serving for millions of daily transactions. Full-lifecycle MLOps across AWS, Azure, and GCP.

TECHNICAL SKILLS

Languages: Python, SQL, Bash

LLM Systems & Agents: LangGraph, LangChain, RAG, MCP server development, tool/function calling, model and prompt routing, agent evaluation and simulation testing, supervised fine-tuning (LoRA/PEFT), guardrails, Hugging Face Transformers, GPT-4, Claude, LLaMA, Mistral, LangSmith, Langfuse

LLM Inference: CTranslate2, int8 quantization, streaming inference, P95 latency optimization, GPU memory management

Machine Learning: PyTorch, TensorFlow, Scikit-learn, XGBoost, LightGBM, deep learning, NLP (BERT), time series forecasting

Speech & Multimodal: Whisper (streaming ASR), speaker diarization (pyannote), NMT (NLLB-200), voice cloning/TTS (XTTS), document understanding (VLM, OCR)

MLOps: MLflow, Kubeflow, Apache Airflow, SageMaker, Evidently AI, model registry, drift detection, CI/CD for ML, automated retraining, LLM evaluation

Data & Infrastructure: Apache Spark, Databricks, Delta Lake, Snowflake, PostgreSQL, Pinecone, FAISS, Chroma, AWS, Azure, GCP, Docker, Kubernetes, GitHub Actions, Terraform

PROFESSIONAL EXPERIENCE

AI/ML Engineer | U.S. Bank *Aug 2024 - Present | Dallas, TX*

- Architected and deployed a RAG-based AI assistant using LangChain, GPT-4, and Pinecone, indexing 500K+ internal documents and cutting research time 70% for 200+ analysts; tuned chunk size, embedding models, and indexing strategy to drop P95 retrieval latency from 800 ms to 180 ms.
- Fine-tuned LLaMA 3 and Mistral 7B with supervised fine-tuning (LoRA/PEFT) on compliance, AML, and KYC documentation, lifting domain question-answering accuracy from 71% to 89% F1.
- Designed end-to-end MLOps pipelines with MLflow, Docker, and GitHub Actions for training, packaging, and deployment to SageMaker endpoints, cutting deployment cycle time from 3 days to under 1 hour; built Spark and Delta Lake feature pipelines with daily refresh for 10+ use cases.
- Built a production customer service chatbot on Azure OpenAI and LangChain, reaching 85% intent-resolution accuracy and deflecting 40% of tier-1 contact-center call volume.
- Implemented real-time monitoring with Evidently AI and Datadog for data, prediction, and concept drift with automated retraining triggers, reducing production incidents 60%.
- Built LLM evaluation harnesses in LangSmith covering hallucination detection, PII safety, prompt injection defense, and answer relevance; defined guardrails and rollout criteria with cross-functional engineering, product, and risk stakeholders.

ML Engineer | Fractal Analytics *Aug 2021 - Jul 2023 | Bangalore, India*

- Built a churn prediction system for a Fortune 100 retail client using XGBoost and neural network ensembles on 10M customer records, achieving 0.91 AUC and reducing attrition 18%.
- Developed fraud detection models with XGBoost, LightGBM, and Spark feature engineering, cutting false positives 25% while holding 96% recall on 2M daily transactions.
- Deployed models on Kubernetes serving infrastructure with horizontal pod autoscaling, improving throughput 3x and cutting serving cost 30%.

- Engineered BERT-based NLP pipelines classifying customer complaints, feedback, and call transcripts for a Fortune 100 client, processing 5M records and enabling automated routing.
- Automated training and scoring workflows on AWS with Apache Airflow, ensuring weekly refresh for 12+ models in production; ran evaluation and drift analysis tracking precision, recall, and stability.

ML Engineer | Cognizant Technology Solutions *Jun 2020 - Aug 2021 | Hyderabad, India*

- Built credit risk scoring models with Logistic Regression, Random Forest, and XGBoost, improving AUC 12% over the legacy rule-based scorecard.
- Deployed ML models on AWS EC2 with Docker and Flask REST APIs supporting batch and real-time scoring for 500K+ daily transactions at under 200 ms P95 latency.
- Developed Python ETL pipelines processing 1M+ customer records daily with Pandas, NumPy, and SQL, cutting data preparation time 45%.

PROJECTS

Mendrift: Autonomous MLOps Incident Response Agent *2026*

Python, LangGraph, LangChain, MCP, Claude, MLflow, Evidently | github.com/suneel190700/mendrift | pypi.org/project/mendrift-mcp

- Built and published mendrift-mcp (PyPI, MIT licensed; listed on the official MCP Registry), an MCP server exposing Evidently drift computation, MLflow registry history and diffing, and model-divergence anomaly detection as typed tools consumable by any MCP client.
- Enforced destructive actions behind a single-use, action-scoped HMAC approval gate implemented in the tool layer rather than the prompt, and verified it adversarially: an LLM instructed to execute a rollback with claimed full authorization and a fabricated token was rejected by constant-time comparison.
- Orchestrated the incident workflow in LangGraph with SQLite checkpointing and a pre-execution interrupt, so an incident survives process death and resumes in a new process after human approval; routed Claude Haiku for classification and verification and Sonnet for a bounded multi-hop diagnosis loop, measured at roughly 3.9K input / 630 output tokens per incident.
- Evaluated on a 19-scenario trajectory harness spanning the full decision space (rollback, retrain, monitor, investigate, graceful degradation, noise, human-declined) against the real graph with a hard zero-ungated-writes assertion, reaching 95% live-model task success; the harness surfaced and drove fixes for three distinct failure classes, including a classifier baited by alert wording and a diagnoser recommending rollback on correlation without causal evidence.

Parlecho: Real-Time Speech Translation with Voice-Cloned Dubbing *2026*

Python, faster-whisper, NLLB-200, XTTS-v2, pyannote, Demucs, CTranslate2, Silero VAD | github.com/suneel190700/parlecho

- Architected a 6-stage dubbing pipeline (Demucs source separation, faster-whisper ASR, pyannote diarization, NLLB-200 translation, XTTS-v2 voice cloning, time-aligned remix) preserving background music and per-speaker voices, plus a streaming mode with resident models and Silero VAD utterance segmentation.
- Measured 0.83 s median / 1.60 s P95 end-to-end dubbing lag on an RTX 4090 under real-time pacing (3.49 s / 6.36 s on Apple M-series CPU), with per-stage latency instrumentation and queue backpressure included in the measurement.
- Benchmarked 5 language pairs on FLEURS: 2.9-5.8% WER on European pairs with whisper-large-v3 and cascade BLEU up to 38.1; decomposed cascade vs oracle BLEU to isolate ASR as the low-resource bottleneck, where upgrading ASR alone recovered +10.4 BLEU on Hindi.
- Quantized NLLB-200 to int8 via CTranslate2 for 38.7% lower peak VRAM (3.72 GB to 2.28 GB) and 4x GPU / 2.7x CPU translation throughput with zero BLEU degradation; validated cross-lingual voice cloning at 0.44-0.58 ECAPA cosine similarity vs a 0.05-0.27 mismatched-speaker baseline.

EDUCATION

Master of Science, Computer Science | University of North Texas, Denton, TX *Aug 2023 - May 2025*