Vineet Srivastava

Nashville, TN, 37203 | +1-312-709-1929 | srivineet93@gmail.com | LinkedIn | GitHub
Patent | Databricks-Certification | Research-Papers/Lectures

SUMMARY

Dynamic and results-oriented AI/ML Specialist with 7.5 years of experience, with deep expertise in Agentic AI (multi-agent systems, A2A, Agent Developer Kit/ADK, MCP), Generative AI (GenAI), and Large Language Models (LLMs). Architect of robust agentic and generative AI platforms for healthcare, leveraging Google Cloud (Vertex AI/Gemini), Neo4j, BigQuery, and multi-agent orchestration to automate and audit clinical pathways, documentation compliance, and RCM processes. Strong background in cloud-based MLOps, RAG, Knowledge Graphs, and Responsible AI. Adept at cross-functional collaboration, deploying scalable solutions, and delivering measurable business value in complex healthcare settings.

SKILLS

- **Programming Languages:** Python, SQL, R, Embedded C, OOPs, Data Structures & Algorithms.
- **Agentic AI:** Agent Developer Kit (ADK), Agent-to-Agent (A2A) Orchestration, Multi-Agent Pipelines, MCP (Model Context Protocol), REST API Microservices
- **Large Language Models (LLMs) & Generative AI**: Transformers (BERT, GPT4, GPT3.5, Gemini-Pro/Flash-2.5, Claude Sonnet 4, MedLM, Text-Embedding-gecko, Text- Embedding-Ada-002, LlaMA2, Flan-T5, Sentence Transformers), Autoencoders- VAEs (for Anomaly detection in Financial Transactions), Intelligent Document Search (Retrieval-Augmented Generation- RAG, Vector Databases- FAISS, Chroma DB, Pinecone, Knowledge Graphs- Neo4J), Evaluation Metrics (ROUGE, BLEU, Recall/Precision, etc.), Ethical Considerations (Biasness, Social Impact, Hallucinations-Detection, Factuality Scores), Prompt Engineering(Zero/Few-Shot and Chain of Thoughts Prompting), Fine-Tuning Techniques (RLHF, SFT, PEFT, DPO/GRPO), Langchain Agents & Tools (LangGraph Agentic AI solutions), Hugging-Face.
- **Platform Engineering, MLOPs/LLMOPs:** Terraform for IAC, PySpark, MySQL, MongoDB, GCP Suite (Vertex AI pipelines, BigQuery, Cloud Run, GCS, Artifacts Registry, Kubeflow, Cloud Build, GKE, Feature Store, Vector Search), AWS Suite (Bedrock, AgentCore, IAM, S3, Lambda, Glue, Athena, SageMaker, RDS, CloudFormation, CloudWatch, API Gateway), Azure Suite (Blob Storage, Data-Factory, Synapse, Cognitive Search, Event-Hubs, WebApps, etc.), Databricks, Kubernetes, Dockers, ETL Pipelining & Orchestration, Kafka, Data Lake, MLOPs (CI/CD with GitHub Actions, LLMOPs, Mlflow).
- **Deep Learning, Computer Vision & NLP**: ANN, CNN Architectures (ResNet, VGG16) for Image Classification & Diagnostics, Model Regularization (Dropout, Batch Normalization), Object Detection (YOLOv4, OpenCV), OCR (Tesseract), Face-Embedding (MTCNN, FaceNet), DeepFace, Text Analytics (NLTK, NER, TF-IDF, Gensim's Word2Vec & FastText), Sequence Processing (RNN, LSTM), Spacy, Spacy-Transformers.
- **Machine Learning**: Exploratory Data Analysis, Feature Engineering, Linear & Logistic Regression, Gradient Descent, Regularization- Ridge and Lasso, Cross Validation, Decision Tree, SVM, Ensemble Techniques (Random Forest, XgBoost), KNN, K-Means, DBSCAN, PCA, Data Drift.
- **Frameworks & Tools**: NumPy, Pandas, PySpark, SHAP, TensorFlow, PyTorch, Langchain, Docker, GitHub Co-Pilot, Flask, FastApi, Streamlit.
- **Healthcare AI**: FHIR/HL7, RCM Analytics, Clinical Data Pipelines, Audit Compliance, Entity Extraction

PROFESSIONAL EXPERIENCE

Amazon Web Services (AWS)

Domain - Multiple

Delivery Consultant- Senior AI/ML Engineer — *Oct 2025 – Present*

- Engaged in building E2E NLP2SQL Agentic-AI solutions for Hospitals & Healthcare Assets location tracking & management, leveraging AWS services like Amazon bedrock, strands agentic framework, *AgentCore* and more.

  Tech: Multi Agentic AI, Claude Sonnet 4, AWS Services (Lambda, Athena, S3, Bedrock, AgentCore)

HCA Healthcare

Domain - Healthcare

ML Engineer 2 (Senior AI/ML Engineer) — *Jan 2024 – Oct 2025*

- Co-architected and delivered an advanced agentic AI audit platform using Google Cloud Run, Vertex AI (Gemini LLMs), Neo4j (MCP), BigQuery, and Agent Developer Kit (ADK), orchestrating multi-agent (A2A) workflows for **clinical guideline compliance and RCM gap analysis**—including LLM-powered extraction, structured-data augmentation, automated knowledge graph evaluation, and explainable reporting at scale.
- Led Terraform-based IaC, CI/CD, error tracing, and autoscaling for all agentic microservices, enabling robust, scalable processing of over 100,000 clinical notes per month.
- Collaborated across clinical, DevOps, and ML teams to optimize prompts, dataflows, and compliance logic—automating audit of 12,000+ patient episodes monthly and achieving a 65% reduction in manual review time and 30% increase in documentation compliance accuracy.
- Spearheaded the development of an AI/LLM-based solution with a team of 3-4 engineers to identify patients eligible for **transcatheter valve therapy (TVT) and trauma care**, analyzing medical histories through the company's enterprise system.
- Engineered the integration of Generative AI using Retrieval-Augmented Generation (RAG) and fine-tuned Large Language Models (LLMs) such as BERT on GCP Vertex AI, for TVT & Trauma pipeline to process both textual records and scanned documents using OCR techniques/Vision based LLMs. Used and trained spacy models for custom token and sentence classification.
- Deployed the solution on Cloud Run, utilizing GitHub Actions and Docker for robust CI/CD across development, QA, and production environments, which led to an increase in the application's precision from 65% to 82%. Enhanced the speed of the pipeline by 40% by deploying the BERT module on vertex-ai pipelines using kubeflow (Nvidia A100 GPUs for accelerated computing) for online serving using custom containers.
- Implemented a modular, scalable **structured entity extraction pipeline** for Trauma Pre-population to highlight business-defined entities in case documents and generate ROSI review links, onboarding new entities via definition review, key-term mapping, module development, integration, and testing. Automated the complete end-to-end pipeline by deploying on GCP cloud run via GitHub Actions to eliminate manual searches, accelerate data abstraction, boost operational efficiency, and drive increased ROI.
- Developed **Hallucination Detection Framework**, combining techniques like semantic similarity scoring, fuzzy matching, and LLM-as-judge based evaluation. This generic framework introduced the team's first hallucination metrics, significantly improving the evaluation speed for Humans in the Loop, reducing evaluation time**,** and enabling faster decision-making across teams.
- Enhanced the hallucination-detection framework with features for resolving complex evidence cases, including exact, similar, semantic, and synthesized evidence matches, and integrated functionality to detect hallucinations, aligning with Responsible AI practices.
- Enhanced **Medical Keyword Identification with Aho-Corasick Algorithm** and built a high-speed bedside procedures identification system for processing

medical histories stored in BigQuery for keyword and category detection. Outperformed traditional regex methods in both speed and accuracy with accuracy of 95%. Integrated LLMs for context evaluation and NegEx detection, achieving identification & processing of complex medical data.

- **Multiple POCs and quick prototyping**: Developing Solutions based on keyword-searches and OCR extraction for extracting textual records and evidences from different medical documents (MIPS Measure AI discovery, Identify missing IP Order for UHC cases, LLM Complex entities extraction for Trauma pre-population using multimodal Gemini LLMs, etc.).

Tech: Agentic AI, LLMs, BERT, Text-Embedding-Gecko, Gemini-Pro-2.0, MedLM, OCR, GCP Vertex AI, Cloud Run, GitHub Actions, Docker, BigQuery, Vertex-AI, Spacy

Dana Farber Cancer Institute (Harvard Medical School Affiliate) for Eastridge Workforce Solutions

Domain - Healthcare

AIOPs Intern — *May 2023 – Aug 2023*

- Analyzed **Patient's Notes to detect different mental disorders** by fine-tuning pre-trained large language models (ALBERT, Clinical BERT, GPT3.5 Turbo, GPT4, LLaMA2) on labeled dataset (300,000+).
- **Implemented Graph RAG** (Retrieval-Augmented Generation pipeline **with Knowledge Graph DB**) for logical contextual retrieval, improving accuracy of diagnoses and relevance of treatment recommendations by 25% on in-house medical data on mental disorders.
- Orchestrated the complete data pipeline using **Databricks Workflow for Patient's Notes and Symptoms** analysis from stage-to-prod using LLMOps and MIflow. Custom fine-tuned BERT model gave better accuracy & F1-score by 8% when compared with existing deployed models.
- Utilized **parameter efficient fine-tuning (PEFT) with low-rank adapters (LoRA)** to reduce model size of previously fine-tuned Flan-T5 based mental health summarization model from 1GB to 30MB while retaining performance.

Tech: Python, Spark, Prompts, Large Language Models (LLMs), Vector Database, Knowledge Graphs, RAG, LLMOps, RLHF, Databricks, Azure

Qualcomm India Pvt. Limited

Domain - IOT/Wireless

Senior Engineer — *Nov 2021 - July 2022*

- Designed & orchestrated an end-to-end pipeline to derive critical metrics and create predictive models to **identify the Bluetooth connection failures** on AWS Cloud. Utilized SageMaker Studio for training and inference of classification models, achieving an AUC Score of 84%.
- Developed pipeline helped identify 1000+ faulty wearable devices and improved efficiency of production by 22%.

Tech: Python, PySpark, Machine Learning, Data Lake Architecture, AWS-S3, IAM, Lambda, Glue, SageMaker, Lambda, Athena, QuickSight, FastText.

Capgemini India Pvt. Limited (Client: Dexcom, Inc, USA)

Domain - Healthcare

Consultant — *Feb 2021 - Oct 2021*

- Architected and built a Python-based analytics framework to perform **root-cause analysis of BLE (GAP/GATT/L2CAP) signal-loss events between Dexcom G5** CGM receivers and transmitters using daily sniffer data.
- Containerized and deployed the framework on Azure WebApps, orchestrated by Azure Pipelines to ingest data from and persist results back to Azure Blob Storage on a daily schedule. Enabled proactive detection of connectivity issues, cutting manual diagnostics and improving device reliability.

Tech: Python, BLE Automation, Azure WebApps, Azure Container Registry, CI/CD pipeline, Dockers, MLOPs

Mirafra Technologies India Pvt. Limited

Domain - IOT

Senior Software Engineer — *Oct 2019 - Feb 2021*

- Implemented end-to-end ML pipeline for **predicting battery status of IoT sensor** in smart building automation, reducing installation costs by 30%, deployed on AWS EC2 using GitHub Actions. Supervised & managed a team of 4 junior engineers in creating the data pipeline and workflow.

Tech: Python, Zigbee, Kafka, MongoDB, Machine Learning, AWS-EC2, S3, ECR, Dockers, CI/CD pipeline, MLOPs.

L&T Technology Services India Pvt. Limited

Domain – IOT/ML

Engineer — *Aug 2016 - Feb 2019*

- Designed and deployed an **LSTM-based time-series model, achieving a 20% boost in indoor localization accuracy** by stabilizing erratic RSSI signals.
- Leveraged LSTM's ability to detect long-term data patterns, significantly reducing localization errors from fluctuating signals.

Tech: Python, Keras, LSTM, Deep Learning, Forecasting, Machine Learning.

EDUCATION

- **University of Illinois at Chicago USA**; Master of Science in Business Analytics; CGPA: 3.8/4.0 — *Aug 2022 - Dec 2023*
- **VIT University Chennai India**; Bachelor of Technology in Electronics & Communication; CGPA: 8.2/10.0 — *May 2012 - May 2016*