

Nuhil Mehdy

San Diego, CA, USA | nuhilmehdy@gmail.com | +1 (619) 346-3232
<https://nuhil.net> | <https://linkedin.com/in/nuhil> | <https://github.com/nuhil>

SUMMARY

With a **Ph.D.** in Computing and **7+ years of experience** across **machine learning, data engineering**, data science, language modeling, and software engineering, I help organizations design and operationalize traditional machine learning solutions and **agentic AI** systems that transform complex data into **actionable, explainable, and valuable intelligence** through production-grade **GenAI software** — without compromising reliability, integrity, or observability.

CORE EXPERTISE

Large Language Model & GenAI

- **Infrastructure & LLMops:** GCP – Vertex AI (Model Garden, Agent Engine, RAG Engine, Model Tuning), Vertex AI Studio, Cloud Run, GKE; LLM platforms & frameworks (OpenAI Platform, Hugging Face, Ollama, LangChain, LangGraph, LlamaIndex, Haystack); API services (FastAPI)
- **Agentic AI Systems:** Multi-agent orchestration (Google ADK, OpenAI Agent Kit, n8n), MCP (Model Context Protocol), A2A (Agent-to-Agent), planning & reasoning (ReAct, CoT), Front-end interfacing (Custom Web Tech)
- **RAG & Enterprise Knowledge base:** Vertex AI RAG Engine, Multimodal RAG Architectures, managed embeddings, hybrid retrieval (vector + metadata filtering), vector backends (Vertex Vector Search, FAISS), document ingestion and grounding pipelines
- **Model Customization & Evaluation:** Vertex AI Model Tuning (supervised fine-tuning, LoRA / QLoRA via PEFT), Quantization, Model Evaluation, RAGAS, TruLens, and observability with Langfuse
- **Inference & Serving Optimization:** Vertex AI Endpoints (traffic splitting, autoscaling), vLLM-backed serving, KV Caching, Triton Inference Server, Ray Serve for high-throughput and low-latency inference

Machine Learning & Statistical Modeling

- **Supervised / Unsupervised:** Linear & Logistic Regression, SVM, Decision Trees, Random Forest, Gradient Boosting (XGBoost, LightGBM), K-Means, PCA
- **Deep Learning:** CNNs, LSTMs, Transformers (BERT, GPT, ViT), Autoencoders
- **Applied Analytics:** Anomaly Detection, Time-Series Forecasting, Clustering, Predictive Maintenance, Hypothesis testing and statistical significance
- **Explainability:** SHAP, LIME, feature importance

MLOps

- **Modeling:** Vertex AI (Dataset, Pipeline, Feature Store, Model Registry, Endpoints, Monitoring, AutoML, BQML)
- **CI/CD & Release:** Cloud Build, Terraform, automated deployments, traffic splitting (A/B) / safe rollouts
- **Monitoring & Governance:** model monitoring, drift/quality checks, auditability and reproducibility practices

Data Engineering

- **Languages & APIs:** Python, SQL, C/C++, JavaScript, Shell, REST APIs, HTML, CSS
- **Modern Data Stack:** BigQuery, GCS (Cloud Storage), Cloud SQL, Cloud Composer (Airflow), Spanner, dbt, Dataflow (Apache Beam), Pub/Sub
- **Data Modeling & Transformation:** Dimensional modeling, schema evolution, and layered design (staging → intermediate → mart) in dbt
- **Governance & Security:** Cloud DLP, Dataplex (Data Profiling, Quality, Policy enforcement, Data Catalog)
- **Visualization & Applications:** Looker, Looker Studio, Streamlit for interactive insight delivery, Tableau

INDUSTRY EXPERIENCE

Staff Data Engineer (AI Engineer) | Dexcom Inc. (May 2023 – Present | San Diego, CA)

- **Develop, deploy & maintain** machine learning and GenAI models for product performance observability, reducing time-to-investigation for reliability issues by ~30% through earlier signal detection, faster root-cause identification, and guided analysis.
- **Architect governed data models & pipelines** (BigQuery, Cloud Composer, dbt, Dataplex) with built-in data quality and governance controls to ensure trusted datasets for analytics and machine learning.
- **Build Agentic AI systems** to enable self-service analytics and guided exploration for internal stakeholders, reducing ad-hoc data and analytics requests by ~40%.
- **Establish observability frameworks** (Vertex AI Monitoring, BigQuery, Looker Studio) to track model and feature drift, ensure data integrity, and reduce operational risk in production systems.
- **Embed ML-driven insights** into operational dashboards, enabling proactive, data-driven decision-making for operations, engineering, and quality leadership.

Machine Learning Engineer | Micron Technology Inc. (June 2021 – March 2023 | Boise, ID, USA)

- **Developed and operated cloud-native machine learning pipelines** with high scalability and availability, reducing model maintenance costs by ~30% through standardized deployment and monitoring practices.
- **Designed and implemented end-to-end ML CI/CD workflows**, cutting model deployment time by ~40% and improving release reliability across production environments.
- **Built production-ready ML solutions** using custom models and Cloud AutoML to support manufacturing stakeholders, increasing overall process efficiency by ~25%.
- **Designed and maintained hybrid ETL/ELT data pipelines** across on-prem and cloud systems, improving data preparation and feature availability speed by ~20%.
- **Collaborated with front-end engineering teams** to integrate ML outputs into user-facing applications, improving usability and driving a ~15% increase in user satisfaction.

Data Science Intern | Micron Technology Inc. (May 2020 – Aug 2020)

- **Developed ML models** to forecast workforce engagement, improving prediction accuracy by ~12%.
- **Built reusable NLP models** for work-life balance perception using hybrid neural architectures on large-scale HR datasets, improving both accuracy and interpretability over baselines.

Software Engineer | Wneeds Ltd. / Freelance | Mobbazaar Inc. (Feb 2012 – Aug 2015 | Dhaka, BD / Remote, CA)

Built RESTful APIs & backend systems (Django, Laravel, Tornado) for high-availability apps (50x faster response).

EDUCATION

Ph.D., Computing — Boise State University, ID, USA

- Focus: Privacy, NLP (Natural Language Processing), Language Modeling | GPA 3.72/4

M.Sc., Computer Science — Lamar University, TX, USA

- Thesis: Deep CNN for Autonomous Vehicle Steering Prediction | GPA 3.64/4

B.Sc., Computer Science & Engineering — Rajshahi University of Engineering & Technology, Bangladesh

RELEVANT PUBLICATIONS

- Transformer-Based Multi-Class Privacy Detection in Natural Language Text – MLNLP 2021
- Sentiment-Aware Privacy Disclosure Framework (Multi I/O Hybrid Neural Network) – PrivateNLP 2020
- FALCON: Anomaly Detection in Industrial Control Systems using Neural Network (CNN, LSTM) – MDPI 2020