PRADEEP VEPADA

pradeep.vepada24@gmail.com , +1 551-362-6948 , Jersey City, NJ, 07306 , linkedin.com/in/pradeep-vepada/ , github.com/PradeepVepada

SUMMARY

AI Engineer with experience building production systems for predictive analytics, data engineering, and generative AI. Skilled in Python, cloud platforms (AWS, GCP), and deploying machine learning models. Proven ability to architect and deliver end-to-end data pipelines and interactive dashboards that drive operational efficiency. Enthusiastic about applying generative AI, LLMs, and agentic systems to solve complex business problems in media and content personalization.

EXPERIENCE

Extern, Pfizer

Consultant

June 2026 – Present

- Developing PixelRAG, a multimodal RAG system for pharmaceutical knowledge retrieval and enterprise query response.
- Engineering hybrid retrieval pipeline with fine-tuning to optimize accuracy and reduce hallucination in biomedical domain.

DXC Technology

4 years 2 months

Project Manager

Sept 2023 - Aug 2024

- Led cross-functional collaboration with 20+ developers, data analysts, and business stakeholders, implementing structured sprint reviews and performance feedback cycles that increased team engagement scores by 80% over two quarters, aligning delivery with product and engineering goals.
- Introduced data-driven project tracking (KPI dashboards, risk heatmaps, velocity reporting) to provide senior leadership with practical delivery insights , driving a 15% increase in on-time project completion rates across the portfolio.

Data Analyst / Data Engineer

July 2020 - Aug 2023

- Engineered end-to-end document processing and data pipelines on AWS (S3 + Redshift + Glue) to ingest, chunk, embed, and transform 1TB+ of sensor data monthly, enabling near-real-time KPI dashboards that reduced manual reporting effort by 40%.
- Built and deployed supervised ML models (Random Forest, Gradient Boosting) in Python/scikit-learn to predict equipment failure and key operational KPIs, leveraging data analysis and statistical methods. This surfaced insights that drove a 15% improvement in operational efficiency and a 10% reduction in unplanned downtime.
- Designed feature engineering pipelines (lag features, rolling statistics, anomaly flags) on structured time-series sensor data , improving model F1-score by 18% over baseline and enabling active maintenance scheduling.
- Migrated on-premise data infrastructure to AWS (Hadoop → EMR + S3) and automated batch ML scoring jobs using Apache Airflow, gaining production experience with cloud platforms and MLOps tools , reducing infrastructure costs by 20% while improving pipeline reliability to 99.5% uptime.
- Developed interactive Power BI and Tableau dashboards surfacing model outputs and trend analysis for business stakeholders , cutting decision-review cycle time by 30% and directly informing quarterly operational strategy.

EDUCATION

Yeshiva University, Katz School of Health and Science

May 2026

Master's in data Analytics and Visualization

Courses: Structural Database Management, Visual Design and Storytelling, Machine Learning for Data Science, Anomaly Detection, Cloud Computing

Anil Neerukonda Institute of Technology & Sciences

Sept 2020

Bachelors in Electronics & Communications Engineering

Courses: VLSI, Electronics and Circuit Theory, Microcontrollers and Architecture, Advanced Data Structures, Operating Systems

PROJECTS

decifur.ai , Biomedical RAG Research Assistant

2024 - Present

- Designed and built a hybrid RAG pipeline combining Neo4j knowledge graphs, FAISS dense retrieval, BM25 sparse search, and reciprocal rank fusion to eliminate siloed retrieval over a 300+ paper biomedical corpus, demonstrating proficiency with structured reasoning approaches and NLP-/ontology-powered layering. Achieved a 35% improvement in answer relevance over single-retriever baselines.
- Integrated UMLS/BioPortal entity normalization and Persistent Conversational Context (PCC) memory into the query pipeline using prompt engineering and in-context learning (ICL) to ground multi-turn scientific Q&A in domain ontologies , reducing entity disambiguation errors by ~40% across evaluation queries.
- Fine-tuned Qwen3.5-27B on 7,700 biomedical chunks (CPT) + 1,515 Q&A pairs (SFT) via Unsloth/TRL on RunPod H100/A100 , to adapt the base model to chemotaxis research language , achieving 87% answer accuracy on held-out domain Q&A vs. 61% for the off-the-shelf model.

AgentIQ , Autonomous ML Pipeline Orchestrator

2024 - Present

- coordinated a 6-stage autonomous ML pipeline (EDA → Feature Engineering → Architecture → Training → Evaluation) using LangGraph for multi-agents with a cross-agent shared memory journal , to eliminate manual handoffs between pipeline stages , reducing end-to-end experiment setup time by 70% vs. a manual baseline.
- Implemented a self-reviewing agent loop (generate → critique → revise) for structured output validation, backed by Firebird persistence and LangSmith for AI observability , to catch and correct suboptimal modeling decisions without human intervention , improving final model evaluation scores by an average of 22% across 10 benchmark datasets.
- Built a FastAPI backend for RESTful API conventions + HTML/JS frontend with MLflow for MLOps and experiment tracking , to make autonomous pipeline runs fully auditable and reproducible , enabling 1-click re-run of any prior experiment with full hyperparameter and metric history preserved.

Accelerator Optimization Copilot

Aug 2024

- Developed a trace-driven simulation platform generating realistic Transformer and CNN memory access patterns across configurable DL accelerator specs , to benchmark cache policy impact under real workload conditions , producing 10,000+ synthetic traces covering 5 distinct accelerator configurations.
- Implemented and benchmarked LRU, LFU, and a custom LLM-driven tensor-reuse scheduler side-by-side , to quantify the advantage of predictive eviction over reactive policies , showing the LLM scheduler achieved 31% higher cache hit rates and 18% lower memory capacity consumption vs. LRU.
- Built an interactive visualization dashboard tracking cache hit rates, memory capacity, and compilation overhead per policy , to surface per-layer optimization opportunities for ML hardware teams , enabling targeted configuration changes that cut simulated inference latency by 24%.

SKILLS & AWARDS

Technical Skills: LLM & AI: RAG (Hybrid, Graph-Augmented), Fine-tuning (SFT/CPT), Agentic Pipelines, GGUF/Quantization, Prompt Engineering, Frameworks & Libraries: LangChain, LangGraph, FastAPI, Unsloth, TRL, HuggingFace Transformers, pandas, numpy, sklearn, TensorFlow, PyTorch, Streamlit, Cloud & MLOps: AWS, GCP, RunPod (H100/A100), Docker, MLflow, LangSmith, Languages: Python, C++, C, SQL, Databases & Search: Neo4j, PostgreSQL, MySQL, SQLite, Pinecone, BM25, FAISS, Tools & Platforms: Power BI, Tableau, Git, Ollama, Deepgram, Twilio