

# PRAJWAL KUMAR

Pittsburgh, PA — +1 (878) 600-3424 — [prajwalk@andrew.cmu.edu](mailto:prajwalk@andrew.cmu.edu) — [LinkedIn](#) — [GitHub](#) — [Website](#)

## SUMMARY

AI/ML Engineer at Infinite Computer Solutions building Agentic AI for automated decision-making. Carnegie Mellon MS in AI grad, IEEE-published; expert in production LLM/RAG, MLOps, and  $O(N)$  neural dynamics via PyTorch, LangChain, and Docker.

## EDUCATION

Carnegie Mellon University	M.S. in Artificial Intelligence Engineering – Information Security	GPA: 3.6 / 4.0 — Dec 2025
Maharshi Dayanand University	Bachelor of Technology in Computer Science & Engineering	GPA: 8.27 / 10 — Jun 2024

## MACHINE LEARNING & AI EXPERIENCE

- Engineered end-to-end Machine Learning pipelines, secure OCR-powered Flask API on Microsoft Azure Computer Vision, processing real-time data streams with Kafka, evaluating models and synthetic data generation with PyTorch, Zeno, and LLMs.
- Trained a high-F1 PyTorch intrusion detection model on GCP Vertex AI Workbench, using PySpark for data engineering. Mitigated class imbalance with a weighted loss MLP and accelerated training with GPU (MPS) and Distributed Data Parallel (DDP).
- Implemented character-level GPT in PyTorch with RoPE; built tokenizer, training loop, and evaluation from scratch. Boosted sequence generation accuracy from <10% to >30% vs. vanilla Transformer using full pretraining + finetuning.
- Designed an end-to-end data system on GCP to predict player value; built PySpark ETL pipeline from GCS to Cloud SQL (PostgreSQL) and trained regression models (SparkML, PyTorch) using Optuna/CrossValidator, deployed via Cloud Run.

## WORK EXPERIENCE

<b>AI/ML Engineer   Infinite Computer Solutions</b>	<b>Jan 2026 – Present   Irving, Texas</b>
<ul style="list-style-type: none"><li>Engineering Agentic AI systems and autonomous workflows to automate complex customer decision-making processes for enterprise clients.</li><li>Designing scalable Generative AI architectures and MLOps pipelines utilizing PyTorch, LangChain, and Docker to ensure production-ready deployment.</li></ul>	
<b>AI/ML Summer Intern   Infinite Computer Solutions</b>	<b>May 2025 – Aug 2025   Irving, Texas</b>
<ul style="list-style-type: none"><li>Built an LLM-powered system with LangChain, crawl4ai, and Playwright to automate workflows and a Neo4j graph from logs, enabling anomaly detection, cutting testing by 90%, and enhancing Verizon's customer journey UX and conversions.</li><li>Developed a RAG system with Neo4j, LangChain, and OpenAI for semantic search on customer journeys logs data; compared Mistral LLMs, optimized PyTorch embeddings, and cut latency by 30% via Streamlit-based interface for Verizon.</li></ul>	
<b>Machine Learning Developer Intern   Qriocity</b>	<b>Jan 2024 – Feb 2024   Chennai, India</b>
<ul style="list-style-type: none"><li>Engineered a deep learning-driven ontology for precise medicine prescription, leveraging TensorFlow and rdflib. Integrated structured medical data to train a neural network, achieving 99% accuracy in predicting optimal medications for diseases.</li><li>Engineered a multi-modal mental health chatbot (text/speech) with a hybrid TensorFlow DNN and KNN model. Predicted emotional VAD scores from TF-IDF vectors, then classified into discrete emotions to provide empathetic user support.</li></ul>	

## ACADEMIC RESEARCH & PROJECTS

<b>Neural Robot Dynamics: Linear-Time Inference and Long-Horizon Stability via Selective State Space Models</b> ( <a href="#">Report</a> )	
<ul style="list-style-type: none"><li>Engineered <math>O(N)</math> neural robot simulator using Mamba-Jamba and Temporal Unrolling; reduced trajectory drift by 50% and matched ground-truth fidelity within -1.37% reward gap.</li></ul>	
<b>Emotion-Aware Multimodal AI Companion - Carnegie Mellon University</b> ( <a href="#">GitHub</a> )	
<ul style="list-style-type: none"><li>Built a multimodal AI companion with speech-to-text (Google API), BERT (text), CNN-based SER (71.2%), and Transformers; improved LLM empathy by 30% via psychoanalysis modules and deployed a Streamlit app for real-time emotion feedback.</li></ul>	
<b>End-to-End Movie Recommender with Kafka, Kubernetes, and A/B Testing - Carnegie Mellon University</b> ( <a href="#">GitHub</a> )	
<ul style="list-style-type: none"><li>Built a scalable KNN-based movie recommender (NDCG@10 = 0.9983, 65K+ QPS) with Dockerized microservices, Kubernetes, CI/CD (Jenkins), and CRON-based retraining; integrated A/B testing and real-time monitoring with Prometheus + Grafana.</li></ul>	
<b>Advancing Image Security through Deep Learning and Cryptography in Healthcare and Industry - IEEE</b> ( <a href="#">Paper</a> )	
<ul style="list-style-type: none"><li>Authored an IEEE conference paper on a novel deep learning-based cryptographic framework using chaotic systems and weight analysis to generate robust, undetectable encryption patterns for securing medical images in IoMT environments.</li></ul>	

## SKILLS

**Languages & Libraries:** Python, SQL, Pandas, NumPy, Scikit-learn, PyTorch, TensorFlow, Apache Spark (PySpark), Streamlit  
**ML & GenAI Tools:** Hugging Face Transformers, LangChain, Mistral AI, OpenAI, MLflow, Vector Databases  
**Cloud & MLOps:** AWS, GCP (Vertex AI, Gemini, BigQuery), Docker, Kubernetes, Jenkins, Prometheus, Grafana, CI/CD, REST  
**Infrastructure & Databases:** Git, Linux, Flask, PostgreSQL, Neo4j, Kafka, Postman