

SANTOSH KANDULA

santoshkandula.dev | linkedin.com/in/santosh-kandula/ | github.com/skandula9273

Applied ML + AI

EDUCATION

University of North Carolina at Chapel Hill, Chapel Hill, NC

Graduation Date: May 2027

B.S. in Computer Science | GPA: 3.7 | Relevant Courses: Deep Learning, NLP, Computer Vision, Algorithms

TECHNICAL SKILLS

Languages: Python, SQL, Java

ML Frameworks: PyTorch, TensorFlow, scikit-learn, Hugging Face Transformers, Ultralytics YOLOv8, OpenPose

LLM & Retrieval: AWS Bedrock, Claude (Sonnet 4.5, Haiku 4.5), pgvector, Jina embeddings, BM25, RAGAS

Data & MLOps: pandas, NumPy, Spark, MLflow, Git, GitHub Actions

Infra & Serving: FastAPI, Flask, Streamlit, Plotly, Docker, GCP Cloud Run, Neon, AWS, Azure, OpenTelemetry

EXPERIENCE

Duke Energy, Machine Learning Intern, Charlotte, NC

May 2026 - August 2026

- **Built** an AWS Bedrock + Claude Sonnet 4.5 predictive LLM that flags PMC errors and delays before schedule slip, with PR-AUC 0.87, median lead time 5 weeks across 9 categories, returned as structured-output JSON
- **Engineered** the RAG ingestion pipeline for Duke Energy PMC's construction-acceleration LLM POC including 1,500+ Primavera P6 XER files in 5 projects plus 35 contractor transcripts, fixed-size chunks under Jina embeddings in pgvector, and retrieval surfaced 72 error-correlation patterns covering 9 error classes
- **Deployed** an on-prem containerized Streamlit interface that pushes live hazard predictions to PMC teams and contractor sites with approximately 10 teams, 25 projects, ~3.34 sessions/user/week

Software Engineer Intern, MADLab

May 2025 - August 2025

- **Replaced** Duke Energy's multi-month manual inventory across 65 acres with 20-30 hour Ultralytics YOLOv8 drone scans over 150+ polygon-annotated SKUs at 86.6% count accuracy, despite heavy stacked-pallet occlusion
- **Drove** Microsoft 365 Copilot adoption at Duke Energy's Optimist Hall, a 500-employee office, where change-management workshops drove 93 active users with 18.6% penetration, with ~2.34 hours saved per user per week, and the program seeded 10 downstream AI demos

UNC Baseball Analytics, Applied Machine Learning Analyst, Chapel Hill, NC

September 2023 - May 2024

- **Mapped** opponent-hitter hot/cold zones for pre-game positioning spanning 500 opposing batters via spray-chart binning on TrackMan data, which dropped opponent BABIP by 0.024 and team runs allowed by 0.7/game
- **Shipped** a live opponent-scouting dashboard on a Flask + Plotly stack with daily TrackMan College and D1 Baseball pulls that replaced an intermediary aggregator and served 48 coaches, players, and staff members daily
- **Designed** an OpenPose 17-keypoint batting-mechanics scoring system for 17 UNC hitters, and joint-angle features fed an sklearn Random Forest ranker that lifted team barrel rate 2.72%

Duke University, Researcher, Durham, NC

May 2022 - August 2022

- **Implemented** a vectorized waveform preprocessing pipeline for noise-handling, null imputation, transformation, and reshape, adopted as the Pratt lab standard across 14 researchers, and config-driven knobs recovered 37% of previously discarded records and cut per-dataset preprocessing from ~4 hours to 32 minutes
- **Trained** a TF Decision Forests gradient-boosted classifier for three-class lightning identification at ~85% per-class precision under a stratified train/test split, deployed for 100+ field analysts to triage incoming strikes

PROJECTS

HoopVector

- **Rebuilt** player tracking from NBA broadcast video, lifting HOTA from 0.301 to 0.525 on SportsMOT basketball-val through ByteTrack, BoT-SORT, a fine-tuned yolov8m detector at 0.987 mAP50, and a training-to-inference resolution fix, each isolated by single-variable ablation
- **Showed** the retrieval encoder's 0.98 recall@1 measured whether a possession matched an augmented copy of itself, not play similarity, scoring 0.51 against a 0.50 random baseline, lifting held-out-game precision@5 to 0.942

SEC RAG Eval

- **Launched** an SEC-filings question-answering service on GCP Cloud Run, FastAPI over Postgres pgvector with 15,192 chunks from 84 10-K/10-Q/8-K filings under text-embedding-3-large, Claude Haiku 4.5 generation, 118 tests, and CI that blocks retrieval regressions
- **Raised** recall@5 from 0.44 to 0.64 on FinanceBench-150 by measuring and retiring five approaches that regressed, including hybrid RRF fusion, a BGE cross-encoder reranker, and dedicated table extraction, then shipping the one that worked, an embedding-model swap, at \$0.009 per query