

Nitheesh Kodarapu

832-786-9794 | nkodarapu@gmail.com | [linkedin.com/in/nitheesh-kodarapu](https://www.linkedin.com/in/nitheesh-kodarapu) | portfolio.nitheeshk.com | github.com/nitheesh-cpu

EDUCATION

Texas A&M University <i>M.S. in Machine Learning & AI (Fast Track)</i> <i>B.S. Computer Science (Honors), Minor in Statistics, GPA: 3.85/4.0</i>	College Station, TX <i>Expected May 2028</i> <i>Expected May 2027</i>
• Honors: Undergraduate Research Scholar (senior thesis on efficient LLM systems, expected December 2026) • Relevant Coursework: Design & Analysis of Algorithms, High Performance Computing, Operating Systems, Distributed Systems & Networking, Machine Learning, Reinforcement Learning, Probability Theory	

EXPERIENCE

IBM <i>Software Engineer Intern, watsonx Orchestrate</i>	San Jose, CA <i>May 2026 – August 2026</i>
• Designed a RAG faithfulness evaluation pipeline using NLI entailment and embedding similarity on a hand-labeled 3,000-query golden dataset; the team adopted the top of 3 benchmarked retrieval backends, reducing hallucinations 35% at 94% citation precision. • Developed asynchronous services syncing 65,000+ GitHub issues into IBM DB2, replacing GitHub's UI with a structured, filterable view and shortening triage time ~30%. • Parallelized a 1,200+ test Playwright and pytest suite with pytest-xdist that gates every release of IBM's flagship agentic AI platform, letting the team ship rapidly without breaking core workflows. • Created a full-stack telemetry dashboard (Next.js, Django REST API) to track client quality metrics across rapid releases.	

Texas A&M University <i>Undergraduate Researcher, Efficient LLM Systems</i>	College Station, TX <i>May 2026 – Present</i>
• Co-authored a paper on LLM quantization and energy efficiency, currently under review at ACM SIGMETRICS. • Wrote the benchmarking software automating the paper's experiments, and ran every test measuring energy by quantization level.	

Waste Management <i>Software Engineer (Part-time) – Promoted from Intern</i>	Houston, TX <i>August 2025 – May 2026</i>
• Automated manifest document extraction with Amazon Bedrock, shrinking processing turnaround from ~1 week to 2–3 days. • Traced and fixed a memory leak with Visual Studio Profiler, lowering RAM 40% in a legacy C# point-of-sale app at 200 scale houses.	
Software Engineer Intern	<i>May 2025 – August 2025</i>
• Interviewed scale house operators about a 20-year-old POS system and redesigned 30+ operator interfaces around their pain points. • Migrated the data pipeline from synchronous to asynchronous calls with lazy loading, decreasing load times 15% at 200 scale houses.	

Cotton Holdings Inc. <i>Software Engineer Intern</i>	Houston, TX <i>June 2024 – July 2024</i>
• Created a map-based dashboard (React, Python) overlaying NOAA weather events on customer locations, shifting sales from waiting for customers to call to proactively contacting those affected. • Enabled the sales team to reach 1,200+ impacted customers within 24 hours of a weather event.	

PROJECTS

TECHNICAL SKILLS

Languages: Python, C++, C#, Java, JavaScript, TypeScript, SQL, C
ML & Data: PyTorch, Pandas, NumPy, Polars, OpenCV, Scikit-learn, LangChain, RAG Pipelines
Systems & Infrastructure: Linux/Unix, Docker, AWS (Bedrock), CI/CD, Multithreading/Concurrency, Win32 API, TCP/Networking
Web & Testing: React, Next.js, Node.js, FastAPI, Django, REST APIs, PostgreSQL, IBM DB2, Playwright, pytest