NITHISH MARNENI

+1 (409) 549-8033 | Email ID | Portfolio | LinkedIn | GitHub | LeetCode

PROFESSIONAL SUMMARY

AI/NLP Engineer with industry and research experience building and deploying ML-driven language solutions. Strong foundation in Python backend development, microservices, and cloud-native systems, with exposure to LLM tooling, evaluation practices, and secure API design. AWS Certified Machine Learning Engineer – Associate.

CORE SKILLS

LLM / RAG: Retrieval-Augmented Generation (RAG), Embeddings, Vector Search, Metadata Filtering, Reranking, Prompt Engineering, Context Management

Frameworks/Tools: LangChain, LangGraph, LlamaIndex, Hugging Face, Transformers

Evaluation/Quality: Offline Eval Datasets, RAG Evaluation (Faithfulness/Relevance)

Backend: Python, REST APIs, FastAPI, Microservices, Async Processing, API Design

Data: SQL, NoSQL Concepts, ETL Basics

Cloud / DevOps: AWS, Docker, CI/CD (GitHub Actions), Logging/Monitoring

Security (LLM): Prompt Injection Awareness, Input/Output Validation, Access Control Concepts

PROFESSIONAL EXPERIENCE

NLP Engineer | Jan 2025 – Present
Doublene | *USA*

- Build and ship AI/ML + NLP capabilities for production applications using Python-based services and modern NLP frameworks.
- Develop Transformer-based NLP workflows (classification, semantic matching, entity/keyword extraction, summarization-style pipelines) with Hugging Face/Transformers.
- Design and deploy FastAPI endpoints for model inference with clean request/response contracts, async processing patterns, and scalable microservice design.
- Improve model reliability by setting up offline evaluation datasets, error analysis loops, and structured testing for model/prompt regressions.
- Implement robust logging/monitoring and debugging workflows for ML services to support stable releases and faster incident resolution.
- Collaborate cross-functionally (product/engineering) to translate requirements into deployable ML features and iterate based on user feedback.

Research Assistant | Mar 2024 – Dec 2024
Lamar University | *Beaumont, TX*

- Conducted research in microservices and cloud computing, supporting scalable architecture design and performance-oriented development.
- Built prototypes and experiments to study distributed system behavior (service communication patterns, reliability, throughput, deployment considerations).
- Designed and implemented fault-tolerant microservice architectures using containerization (Docker) and orchestration patterns for high availability.
- Analyzed performance bottlenecks in distributed systems through load testing and profiling, achieving measurable throughput improvements.
- Developed automated testing frameworks for microservice integration testing, reducing regression bugs and improving deployment confidence.
- Collaborated with faculty advisors to publish research findings and present technical results at departmental seminars.
- Produced technical documentation and research deliverables with reproducible experiment setups.

Software Developer   **Feb 2022 – Jun 2023**
Techecy   *India*

- Developed and maintained backend microservices for a large-scale messaging/platform ecosystem, focusing on maintainability and performance.
- Built and integrated REST services with structured debugging practices, improving service stability in production environments.
- Implemented message queue systems and event-driven architectures to handle high-volume asynchronous processing workloads.
- Optimized database queries and implemented caching strategies, reducing API response times by significant margins.
- Designed and deployed CI/CD pipelines using Jenkins and Git workflows, streamlining the release process and reducing deployment errors.
- Wrote comprehensive unit and integration tests to ensure code quality and maintain high test coverage across services.
- Participated in code reviews and mentored junior developers on best practices for clean code and software design patterns.
- Partnered with teams to troubleshoot issues, support releases, and implement engineering improvements across the platform.

FEATURED PROJECT

LLM-Powered Knowledge Assistant (RAG + Evaluation + Secure APIs)

- Built an end-to-end RAG pipeline: document ingestion, chunking, embeddings, vector search, and grounded generation with citations.
- Implemented retrieval quality improvements using metadata filtering, reranking, and query refinement patterns to reduce irrelevant context.
- Added offline evaluation using curated test queries and scored outputs for faithfulness/relevance, enabling repeatable regression checks.
- Delivered the system as a FastAPI service with structured logging, basic guardrails (input/output validation), and Dockerized deployment.
- Designed the architecture to be cloud-ready on AWS, aligning with production expectations (CI/CD-friendly structure, monitoring hooks).

CERTIFICATION

AWS Certified Machine Learning Engineer – Associate

EDUCATION

Master of Science in Computer Science   **Jul 2023 – May 2025**
Lamar University, Beaumont, TX, USA

Bachelor of Technology in Electronics and Communication Engineering **May 2018 – Jul 2022**
JNTU, Hyderabad, India

AWARDS & RECOGNITION

Outstanding Graduate Student Award, Lamar University   **(2024–2025)**
Awarded for exceptional academic performance and research in microservices and cloud computing.

Outstanding Individual Contribution Award, Techecy   **(2022–2023)**
Recognized for impactful microservice development and improving large-scale messaging platform performance.