Noshitha Padma Pratyusha Juttu

San Jose, California

noshithajuttu@gmail.com | +1 4134724306 | /in/noshitha-juttu | github.com/Noshitha | hf.co/Noshitha98

EDUCATION

University of Massachusetts, Amherst — **Massachusetts, USA**
M.S. in Computer Science - GPA: 3.9/4.0 — *Jan 2024 - Dec 2025*

VNR Vignana Jyothi Institute of Engineering & Technology — **Hyderabad, India**
B.Tech. in Information Technology - GPA: 9.1/10.0 — *July 2017 - July 2021*

TECHNICAL SKILLS :

LLM / ML Systems: PyTorch, Hugging Face Transformers, ONNX Runtime, CoreML, quantization- PTQ/QAT, INT8/FP16 inference,vLLM
AI Systems: LangChain, LangGraph, multi-agent orchestration, RAG pipelines, ReAct agents, RLHF (GRPO,DPO)
Retrieval / Representation: FAISS, Milvus, Sentence-Transformers, Neo4j, LlamaParser
Programming: Python, SQL, Scala, Java
Distributed Data / Infra: Spark, Databricks, AWS (S3, Redshift, Athena, SageMaker), Airflow, Docker, Kubernetes, Terraform
Evaluation: BLEU, ROUGE, COMET, BLEURT, ablation studies, regression testing

PUBLICATIONS

When Consensus Becomes Compliance: Measuring Sycophancy in Multi-Agent Language Model Interactions — *June 2026*
(ACL 2026 Student Research Workshop, Under Review)

- Introduced the Conditional Infection metric to quantify interaction-driven epistemic regression in multi-agent LLM debates.

Text to Trust: Evaluating Fine-Tuning and LoRA Trade-offs in Language Models for Unfair Terms of Service Detection — *Oct 2025*
arXiv preprint (arXiv:2510.22531)

- Systematic evaluation of full fine-tuning and parameter-efficient LoRA adaptations focused on clause-level classification and risk flagging in real-world legal contracts.

Development of an AI-Based Chatbot Using Deep Neural Networks — *Oct 2021*
International Conference on Intelligent Vision and Computing 2021

- Authored and presented research on chatbot development using Bag of Words, DNNs, and batch gradient descent

EXPERIENCE

UMass BioNLP Lab | *Grad Student Researcher* (Advisor: ***Prof. Hong Yu***) — *Sep 2025 – Jan 2026*

- Built a training-free multi-agent LLM framework for SDOH prediction from clinical text, using reward-guided iterative refinement to improve classification without fine-tuning.
- Designed an inference-time multi-agent reasoning pipeline that ranks candidate outputs for correctness and consistency, improving prediction stability under limited supervision without additional model training.
- Implemented a lightweight memory module to reuse high-reward reasoning patterns across predictions, reducing redundant inference-time exploration.
- Evaluated the system on alcohol-use classification from clinical notes and analyzed trade-offs in accuracy, macro-F1, robustness, and reasoning stability.

Adobe | *UMass Industry Externship* (Advisor: ***Prof. Andrew McCallum & Franck Dernoncourt***) — *Jan 2025 – May 2025*

- Engineered an on-device inference optimization pipeline for MarianMT-based neural machine translation using vocabulary pruning, transformer slimming, and PTQ/QAT quantization for efficient edge deployment.
- Built a production-oriented **PyTorch-to-CoreML/ONNX Runtime** deployment workflow enabling **INT8/FP16** export and efficient cross-platform inference.
- Reduced model size from **75M to 23M parameters**, improved decoding throughput by **~20%**, and achieved **~65–70% smaller** quantized ONNX exports with limited quality loss.
- Evaluated accuracy-latency-size trade-offs using **BLEU, TER, chrF, METEOR, COMET, BERTScore, and BLEURT** to define deployment thresholds.

Deloitte USI | *AI & Data Engineering Analyst* — *Sept 2021 – Jan 2024*

- Architected and operated large-scale Spark/Databricks batch and near-real-time data pipelines across multi-terabyte datasets, processing **1M+ records/day** for production analytics and downstream enterprise reporting.
- Designed and maintained **REST/SOAP API ingestion pipelines**, extracting structured JSON/XML data into governed **S3 staging layers** for downstream ETL processing.
- Migrated legacy Informatica workflows into distributed **Databricks** pipelines, reducing batch runtimes by **30%**
- Led root-cause analysis for critical production pipeline failures, redesigning workflows to eliminate recurring defects, reduce compute/memory overhead, and improve operational reliability; recognized with **two Spot Awards** and client appreciation.

PROJECTS

RAG-based Research Copilot

- Designed and implemented an end-to-end multi-agent RAG system using LangGraph and MCP-based tool interfaces for automated paper retrieval from arXiv, abstractive summarization, and semantic topic clustering.

BriefCheck: Legal Citation Verification and Trust Review

- Built an AI-assisted legal review system that verifies case citations and legal claims against supporting sources, surfaces unsupported or outdated authority, and assigns trust-oriented review signals for downstream drafting.
- Framed the workflow as a verification and triage layer for legal writing, combining model-based extraction with external grounding and structured evidence checks.