Koushik Reddy Parukola

Bloomington, IN (703)-789-1618 parukolakoushik@gmail.com Website LinkedIn GitHub

SUMMARY

AI Engineer and Research Associate with expertise in building AI-driven solutions for extracting insights from unstructured data, optimizing large language models, and enhancing information retrieval. Experienced in developing as well as fine-tuning LLMs to improve efficiency and scalability. Skilled in transforming complex text data into structured knowledge, enabling better decision-making and automation of AI models. Passionate about leveraging AI to drive innovation, streamline workflows, and enhance data-driven strategies.

Work & Research Experience

Machine Learning Engineer, AI Consultants July 2024 – Present

- Built agentic AI workflows for insurance firms and fintech platforms, automating high-friction processes like chargeback handling, claim validation, and KYC audits.
- Used Gemini 2.5 Pro and GPT-4 for document intelligence tasks (extraction and interpretation) across invoices, transaction logs, and compliance forms. And Mistral for French-language document parsing
- Orchestrated stateful multi-step workflows using LangGraph, supporting complex flows like evidence preparation, user query resolution, and document verification.
- Integrated retrieval-augmented generation (RAG) for grounding model outputs in proprietary policy rules, dispute documentation, and regional compliance guidelines.
- Deployed serverless infrastructure using AWS Lambda (orchestration), Amazon S3 (secure storage), and Bedrock (LLM integration - GPT-4 and Mistral).

Research Associate - NLP Lab, Indiana University Aug 2023 – Present

- Reduced high-dimensional GPT embeddings (e.g., 3072-d) using PCA, t-SNE, and Fisher's Discriminant to analyze over 10K+ token vectors, identifying semantic drift and class-based clustering.
- Investigated cross-attention alignment patterns between source and target sequences to understand how attention heads contribute to semantic grounding and structural consistency.
- Fine-tuned llama using LoRA and evaluated the trade-offs between quantized and unquantized settings, particularly in biomedical domains where high precision is essential.
- Analyzed how transformer layers drift under LoRA-based tuning and quantization, using probing classifiers and interpretability tools to assess feature stability.
- Explored attention-based interpretability methods like attention rollout, to analyze how transformers capture logic, syntax, and semantic roles—informing design of lightweight models for tasks like entity linking.
- Building an agentic framework where AI agents autonomously handle linguistic decomposition and generate knowledge graphs from LLM outputs using RAG-style retrieval pipelines.

NovelTronix - Junior Machine Learning Engineer Jan 2021 – July 2022

- Designed CNN-based vision models for traffic sign detection and distance estimation, achieving 92% accuracy on GTSRB and adapting models to Indian traffic signs for ADAS simulation.
- Developed ARIMA and LSTM time-series models to forecast ride demand, reducing MSE by 20% and optimizing fleet allocation during peak hours.
- Engineered GPS telemetry features (speed, stop events) to support trip prediction and operational analytics.
- Collaborated with product and ops teams to define ML success metrics, improving route optimization KPIs by 4–6% and reducing model deployment time by 10% in real-time vehicle dispatch systems.

Talks & Publications

- Damir Cavar, Koushik Reddy Parukola "Text Similarity Using Classical Word Embeddings in Quantum Systems," IEEE - International Conference on Acoustics, Speech, and Signal Processing 2025, Hyb, India.
- Damir Cavar, Koushik Reddy Parukola, Shane Sparks, Talk on Using "Word Embeddings in Quantum NLP Systems", Midwest Speech and Language Days 2025, University of Notre Dame, Indiana.

- Koushik Reddy Parukola, Shane Sparks, Presented in the Quantum NLP grant (funded by the CQT, as an NSF Center, PI Damir Cavar), Fall 2024 Meeting CQT at Indiana University.
- Damir Cavar, Koushik Reddy Parukola "Automated Knowledge Graphs with Ontologies towards better information retrieval," Neurosymbolic Learning and Reasoning 2025, UC Santa Cruiz (Under Review).
- Traffic Sign Classification - Capstone Research Project - 'IJISRT' Volume 6 - Issue 6 - June 2021.

Projects

Grpah RAG for Medical Document Parseing Jan 2024 – Apr 2024

- Developed Graph RAG pipelines for medical document parsing, integrating knowledge graphs with LLMs to improve retrieval and reasoning in drug-disease relationships.
- Extracted structured knowledge from NCBI biomedical literature and UMLS ontologies, focusing on drug interactions, disease progression, and treatment efficacy.
- Used Stanza and SpaCy for text preprocessing, including NER and dependency parsing, and Fine-tuned LLaMA with LoRA to generate RDF-based knowledge graphs for drug-disease relationship extraction.
- Evaluated knowledge graphs with triplet accuracy, semantic coherence, and graph connectivity, ensuring high-quality extractions for medical applications.
- Compared structured outputs with gold-standard datasets to assess graph-based reasoning in clinical and pharmacological contexts.

Cancer Drug Synergy Prediction, Bioinformatics ML Jan 2023 – July 2023

- Developed neural network models to predict cancer drug synergy, assessing combined drug effects across 60 human cancer cell lines from NCI ALMANAC, achieving an AUC of 90.81.
- Implemented Graph Convolutional Networks (GCNNs) to learn molecular graph representations of drugs and cell line features, capturing structural and relational patterns to enhance predictive performance.
- Conducted comparative analysis with feedforward networks, autoencoders, and extremely randomized trees (ERTs), outperforming traditional baselines across synergy score prediction tasks.
- Integrated multi-modal features including drug fingerprints, target genes, and gene expression profiles, optimizing models with ReLU activations and deep feature embeddings.

Technical Skills

• **Programming Languages:** Python, Rust, Java, SQL, R.
• **AI/ML:** AWS: Sagemaker, Bedrock, S3, Textract. Neural Models: CNNs, RNNs, LSTMs. Reinforcement Learning, GANs, Vector Models, Knowledge Graphs, GraphRag/RAG models, Transformers, LoRA, LLMs.
• **Tools:** PyTorch, TensorFlow, MLflow, Faiss Vector Database, Neo4j, Weights & Biases, Docker, MLaaS.
• **Visualization tools:** Tableau, Power BI, Unity Engine & Paraview (3D Simulations).

Awards, Certifications & Community Involvement

• **Vice President** - App Development Club @ New Horizon College of Engineering(NHCE) Aug 2019 - May 2021
• **Core Committee - Quantum NLP-AI conference 2025**, Hosted by Indiana University (organized in collaboration with the ACM, Special Group of Interest in AI). qnlp.ai
• **IBM Quantum Challenge 2024** – Top 11% - Q-Simulators, Quantum ML, VQEs, Parameterized Circuits.
• Member of IEEE, ACM, and ACM Special Interest Group on Artificial Intelligence (SIGAI).
• Certified AWS AI Practitioner - 2025

Education

Indiana University
Masters of Science in Data Science Aug 2022 – May 2024
Coursework: Applying ML techniques in NLP, Applied Machine Learning, ML in Bioinformatics, and Data Mining.

New Horizon College of Engineering
Bachelor of Engineering in Computer Science July 2017 – Aug 2021
Coursework: Operating Systems, Compiler Design, Data Structures, Machine Learning, Case-Based AI.