SAI PUNITH KOLLA
Boston, MA | kolla.sa@northeastern.edu | **LinkedIn** | **Ph:** 8574215939

EDUCATION:

- **Master of Science in Artificial Intelligence**   **December 2026 (Expected)**
  Khoury College of Computer Sciences, Northeastern University, Boston, MA.
  Coursework: Machine Learning, Natural Language Processing, Algorithms, Foundation of Artificial Intelligence, Programming Design Paradigm.
- **Bachelor of Technology in ECE with spec. in Sensors and Wearable Technology**   **June 2022**
  Vellore Institute of Technology, Vellore, India.
  Coursework: Artificial Intelligence with Python, Neural Networks and Deep Learning, Data Analysis, Data Structures and Algorithms, Calculus, Applied Linear Algebra, Essentials of Machine Learning, Wearable Technology and IOT.

SKILLS:

Languages: Python (EDA, ML, API development), Java, Shell Scripting, R (statistical modeling), SQL
AI: Deep Learning, Reinforcement Learning, Natural Language Processing (NLP), Machine Learning, Prompt Engineering, Transformer Architectures, Sentence-Level Modeling, Retrieval-Augmented Generation (RAG), Model Evaluation.
Frameworks, Libraries: Flask, PyTorch, TensorFlow, Keras, LangChain, LangGraph, Hugging Face Transformers, Sentence Embedding Models, Variational Autoencoders (VAE), OpenCV, FastAPI, RESTful APIs, Postman, Pygame
Data & Cloud: MySQL, Snowflake, Power BI, Tableau, Informatica IICS, Hadoop, Google Cloud Platform , FAISS, CI/CD, Git Version Control.
Big Data & Analytics Tools: Apache Spark, Databricks, Excel, Data Cleaning, Feature Engineering, Statistical Analysis
Software & Collaboration Tools: JIRA, Confluence, Twilio APIs, Docker (Basic), Agile Development
Conceptual Skills: Data Preprocessing, Data Integration, Representation Learning, Semantic Search, Workflow Automation, Error Prediction Systems, Data Governance, Optimization Algorithms.

WORK EXPERIENCE:

Python Developer – Level 2 | IICS Developer -Wipro Technologies, Bengaluru, India   **April 2022 – October 2024**

- Developed and deployed an internal **Flask-based LSTM-powered error prediction system** trained on historical logs and enriched with NLP-driven rule-based matching to predict and classify recurring issues.
- Built a **knowledge base for automated error resolution** and integrated it with monitoring dashboards, reducing manual investigation time by **40%** and improving diagnostic accuracy.
- Built and managed **REST APIs** connected to SQL databases to enable unified data access, incident tracking, and automated reporting across teams; led the **complete API lifecycle**, including Postman-based endpoint validation, Git version control, and seamless multi-environment (Dev/UAT/Prod) deployments.
- Automated recurring workflows, file operations, and housekeeping scripts using **Python and Shell scripting**, saving ~20 developer hours weekly. Collaborated with infrastructure teams for **secure agent configuration, DR simulations, certificate renewals, and server health monitoring**, ensuring platform reliability and security compliance.
- Designed and implemented **parameterized ETL workflows and reusable mappings** in Informatica IICS, improving data integration efficiency and scalability while reducing pipeline redundancy and maintenance overhead.
- Orchestrated ETL job deployments and migration pipelines using **Bamboo CI/CD**, enabling faster, more consistent releases. Built **interactive Tableau dashboards** visualizing key KPIs such as error frequency, load times, and ETL performance, providing actionable insights for proactive decision-making.
- Participated in **agile sprint planning, code reviews, and process documentation**, promoting traceable workflows and faster onboarding for new engineers. Collaborated cross-functionally with QA, DevOps, and business teams to ensure alignment, transparency, and continuous delivery of production-ready solutions.

PROJECTS:

Sentence-Level Transformer for Next-Sentence Prediction *(Python, Transformer, Latent Modeling, NLP)*   **Aug 2025 – Dec 2025**

- Designed a **sentence-level language model** that performs attention over complete sentences (unit tokens) instead of individual words, shifting from next-token prediction to **next-sentence prediction**.
- Implemented a **two-stage architecture** combining a Sentence-VAE for latent sentence representation with a latent-conditioned Transformer decoder to model discourse-level coherence.
- Evaluated the approach on **WikiText-2**, analyzing training stability, latent quality, and generative coherence compared to token-level baselines.

Royal Rumble Wrestling Simulation *(Python, PPO, Simulated Annealing, Genetic Algorithms, Pygame)*   **Feb 2025 – Apr 2025**

- Built a **multi-agent reinforcement learning environment** simulating adaptive combat behaviors using **PPO**, fitness-based initiator selection, and dynamic policy updates. Combined **Genetic Algorithms** for evolving wrestler traits with **Simulated Annealing** for probabilistic exploration, visualized through a **turn-based Pygame interface**.

Image Captioning System *(Python, TensorFlow, Keras, ResNet50, LSTM, Flask)*   **Jan 2022 – May 2022**

- Implemented an **image captioning model** combining a **ResNet50 CNN encoder** for visual feature extraction and an **LSTM decoder** for sequence generation.
- Built and deployed a **Flask web interface** allowing users to upload images and view generated captions in real time, achieving contextually relevant and grammatically coherent descriptions.