CHENCHAIAH MEKALATHURU

+1 (605) 585-4711 | mekalathuru.chenchaiah@gmail.com | United States | linkedin.com/in/iamchenchu/ | github.com/iamchenchu

SKILLS

- **Deep Learning:** CNNs, RNNs, LSTMs, VANs, GANs, MLPs, Autoencoders, Deep Q-Networks, DDPG, Transformers, Federated Learnig, Fine Tuning, Image Captioning
- **NLP:** Text Preprocessing,, Text Classification, NER, Text Summarization, NLTK/SpaCy, Language Modeling (Seq2Seq, Transformers), Word2Vec, Hugging Face, Python NLTK, SpaCy, Bag of words
- **Other Tools & Libraries:** Pytorch, Keras, Tensorflow, Scikit-learn, OpenCV, Databases (MySQL/MongoDB), BI (PowerBI/Tableau), Web Frameworks (Flask/Django), AWS, Microsoft Azure, Docker, Kubernetes, REST APIs, Google Cloud Platform (GCP), Hugging face, Hadoop, Apache Kafka, Databrics, command line, Apache Spark, Ollama, LlamaIndex, LangChain, LangGraph, LangSmith, LangFlow, Snowflake, Streamlit, Databricks, gRPC, MongoDB, DynamoDB, FastAPI, Airflow, Terraform, Kubeflow, ETL, BigQuery
- **Programming :** Python, Java, R, C/C++, JavaScript, CUDA, Julia, Rust, XML, Go, Scala, HTML/CSS, React, AngularJS, Node.js
- **Machine Learning:** Dimensionality Reduction Algorithms, Matplotlib, Seaborn, Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), Pandas, K-Nearest Neighbors (KNN), Naive Bayes, K-Means Clustering, Gradient Boosting, AdaBoosting, NumPy, RL, GPU, Active Learning

EDUCATION

University of South Dakota **August 2023 - December 2024**
Master's, Artificial Intelligence *GPA: 4*

- Machine Learning, Distributed Systems, Quantum Computing, Computer Vision, Applied Mathematics, Data Science, Recommendation Systems, High Performance Computing (HPC), Containerization, Design Patterns, Big Data, Data Modeling, Vertex AI, Data Pipeline

Lovely Professional University **August 2015 - May 2019**
Bachelor's, Computer Science *GPA: 3.5*

- Web Development, Statistical Reasoning, Android Development, Dynamic Programming, Operating Systems, Mathematics, SQL, Statistics, Linux, Data Structures, Version Control, CI/CD, Cloud Computing, Data Analytics, Microservices

PROFESSIONAL EXPERIENCE

University of South Dakota **Vermillion, SD, USA**
Research Assistant *August 2023 - December 2024*

- Worked on Liquid Neural Networks, gained knowledge on multiple research papers (CTRNN, Neural ODE, NPCs) to build models on LNN, also on state-of-the-art architectures in open source LLM's, NLP, Transformers, Fraud Detection, Generative Al which is resulting in an 80% accuracy in model performance at least.
- Contributed to two projects involving Liquid Neural Networks, focusing on building, training, and evaluating models while working on diverse tasks related to mathematical software development. Developed a model for detecting cancer in images of various organs across different modalities, achieving superior results with limited resources. Authored a research paper based on this work and submitted it to CVPR 2025, currently awaiting the review outcome.
- Enhanced understanding of machine learning and AI concepts among 10+ undergraduate students by guiding them in resolving their doubts and queries, while efficiently managing lab operations using R for practical demonstrations.
- Achieved high user satisfaction by using LLM-driven personalisation, dynamically adjusting property descriptions based on buyer preferences. This demonstrated the impact of contextual augmentation and improved the overall user experience through personalised content generation.

Byju's **Hyderabad, Telangana, India**
Senior Machine Learning Engineer *November 2019 - July 2023*

- Developed BYJU's Q&A search system using neural search techniques by fine-tuning a pre-trained BERT model on an in-house Q/A dataset with contrastive loss for similarity, building an ML pipeline with ElasticSearch and Annoy Indexes, and exposing Flask APIs for similar questions, achieving an SSR of 80% with 2 million monthly hits and boosting revenue generation to $600k/month in sales leads.
- Trending Content: built universal recommendations using a statistical model, Used chi-square to detect trends in chapters and subtopics using 10M interaction of data and deployed on production for all Byju's users.
- Elevated AI capabilities for educational applications to 13% accuracy on GSM8k datasets, meeting business objectives by fine-tuning LLaMA and Falcon LLMs.
- Scalability and Agile Development: Managed product feature launches that resulted in a 40% revenue increase within three months, reflecting your skills in scaling solutions critical in AI system deployment and scaling AI/ML pipelines.
- Built a cGAN-based model for watermark removal from solution images, aiding Byjus dB data ingestion. Utilized OpenCV, U-Net with pre-trained ResNet-101, PatchCNN.
- Enhanced classroom engagement tracking by accomplishing a mAP50-95 score of 0.922 as measured by real-time monitoring accuracy in classrooms by fine-tuning YOLOv5 models on annotated data to detect upper body objects effectively.

- Developed revenue forecasting models using ARIMA, Prophet, and LSTM, leveraging regional appointment and demo data to predict trends and drive strategic decisions as part of risk management.

Applied Artificial Intelligence Club
President *August 2024 - December 2024*

- Elevated club membership by 50% compared to the prior year by introducing agendas that focused on real-world applications of AI and coding workshops, equipped members with skills necessary for successful careers in the tech industry.
- Led and organized weekly interactive sessions on AI topics, including building Large Language Models (LLMs) and coding practical AI models from scratch.
- Collaborated cross-functionally with faculty advisors and leading industry experts to design a series of 10 engaging workshops, enhancing members' understanding of AI research and career opportunities in advanced technology sectors.
- Developed and delivered technical lectures on AI concepts such as GPT models, transformers, and neural networks, fostering peer learning and hands-on coding experiences.
- Implemented agile methodologies by utilizing Scrum to streamline project management of club activities and workshops, ensuring timely delivery and high engagement.

PROJECTS

AI-Powered Real Estate Matching Application [PyTorch, Python, Seaborn, OpenAI, LLMs, DevOps]

- Designed a robust architecture integrating GPT-4 for automatic generation of property listings and descriptions, achieving an average turnaround time of under 15 seconds per listing while ensuring high-quality output.
- Integrated ChromaDB as the vector database, enabling efficient storage and retrieval of property embeddings. This setup facilitated fast and accurate semantic search using user preferences, showcasing advanced vector-based querying techniques.
- Implemented an NLP-based preference parser to convert user inputs into structured queries, allowing precise, real-time matching of buyer requirements to listings. This enhanced the recommendation system by utilizing natural language understanding (NLU) capabilities.
- Achieved high user satisfaction by using LLM-driven personalisation, dynamically adjusting property descriptions based on buyer preferences. This demonstrated the impact of contextual augmentation and improved the overall user experience through personalised content generation.

GPT From Scratch – Personal Assistant and Classifier [Python, PyTorch]

- Developed end to end GPT model with 124 million parameters from scratch without using pre-defined code blocks, utilizing 12 Transformer decoder blocks and 24 multi-head self-attention (MHA) layers per block, integrated with GPT-2 pre-trained weights for enhanced performance.
- Fine-tuned the model for classification and instruction-based tasks, including email classification and active passive sentence transformations, leveraging domain specific datasets for improved task accuracy.
- Designed and built an efficient tokenization pipeline of Bit Pair Encoder (BPE) to preprocess raw text into numerical representations, optimizing input for Transformer based architectures.
- Trained the model on large-scale datasets, employing advanced techniques such as multi-head self-attention (MHA) and layer normalization to enhance computational efficiency and model performance.
- Evaluated the fine-tuned model using Llama 3 (8B) for benchmarking, ensuring robust performance in classification and question-answering tasks.

RAG-Deepseek (Conversational AI for PDFs using LangChain, FAISS)

- Developed a Retrieval-Augmented Generation (RAG) system using OLamma Deepseek-r1:7b, achieving 95% precision in document retrieval and reducing query latency by 30% through optimized pipelines and vector embedding techniques.
- Built and deployed a PDF chat reader application using Streamlit and LangChain, enabling real-time conversational AI interactions with unstructured PDF data, achieving an F1 score of 92% and sub-second query response times.
- Integrated ChromaDB and FAISS for efficient storage and retrieval of embeddings, scaling the RAG application to process over 1 million records with 99.9% uptime in a Streamlit-based environment.
- Deployed applications in Streamlit with robust error handling, increasing reliability by 20% and ensuring seamless accessibility for technical and non-technical users.