CHENCHAIAH MEKALATHURU

605-585-4711 | San Francisco, CA | Email | LinkedIn | GitHub | Portfolio

SUMMARY

Generative AI Engineer with 5.5 years of experience developing production-grade LLM and RAG systems using Hugging Face, Azure ML, and PyTorch. Experienced in building AI-powered applications for automation, personalization, and document intelligence.

TECHNICAL SKILLS

- **ML Algorithms:** Regression, Classification (SVM, XGBoost, LightGBM), Clustering (K-Means, DBSCAN), Recommendation (DeepFM)
- **Deep Learning Models:** CNNs, RNNs, LSTMs, GANs, Autoencoders, Attention Mechanisms **|Languages:** Python, SQL, C# (basic)
- **NLP & Generative AI:** Transformers, RAG (LangChain, FAISS, ChromaDB), Prompt Engineering.
- **ML/DL Frameworks:** PyTorch, TensorFlow, Hugging Face, Keras, Scikit-learn, Lightning, Accelerate
- **MLOps & Deployment:** MLflow, Kubeflow, Airflow, Docker, Kubernetes, FastAPI, TorchServe, CI/CD, Terraform, BentoML
- **Data Engineering:** Apache Spark, PySpark, Kafka, dbt, Data Pipelines, Data Ingestion, Feature Engineering
- **Cloud Platforms:** AWS (S3, Lambda, SageMaker), GCP (Vertex AI, BigQuery), Azure (Synapse, ML Studio)
- **Databases & Storage:** SQL, Snowflake, BigQuery, Redshift, MongoDB, FAISS, Pinecone, ChromaDB, Embeddings (SBERT, OpenAI)
- **DevOps & Tools:** GitHub, GitHub Actions, CircleCI, Tableau, Jira, Confluence **| Cloud AI Services:** Azure OpenAI Service, Azure Cognitive Services **| Business Automation:** Appian (beginner), Power Automate (exploring), BPM, Process Orchestration

PROFESSIONAL EXPERIENCE

ML RESEARCH ASSISTANT: *University Of South Dakota – Vermillion, Sd* **Aug 2023 - Dec 2024**

- Scaled LNNs from research to production; delivered 25%+ system efficiency gains in multimodal ML pipelines.
- Mentored 10+ students in ML engineering and reproducibility; improved code quality and project delivery speed by 40%.
- Built Liquid Neural Networks (LNNs) with CTRNNs and Neural ODEs; achieved >80% accuracy on low-resource compute environments.
- Led LNN model development for multimodal cancer detection; boosted diagnostic accuracy by 18% using X-ray, MRI, and tabular data.
- Benchmarked LLMs and Transformers for NLP, fraud, and GenAI; improved system efficiency by 25% across applied research projects.

SENIOR MACHINE LEARNING ENGINEER: *Byju's - Hyderabad, India* **Jul 2021 - Jun2023**

- Launched 3 AI systems in 3 months using Agile workflows, accelerating deployment cycles, driving a 40% increase in company revenue.
- Fine-tuned BERT & LLaMA via Hugging Face for RAG pipelines; scaled semantic search and decision systems for 2M+ user sessions.
- Automated CI/CD for ML training, ETL, and deployment; cut manual ops by 40% and improved pipeline reliability and system uptime.
- Built BERT + Annoy journey optimizer for personalization; boosted $600K/month revenue and 80% SSR across 2M+ user queries.
- Supervised 6 ML engineers and contributed to compiler-level model tuning to optimize performance for production inference pipelines.
- Defined ML roadmap with product/infra; led scalable LLM personalization system design aligned with revenue and engagement goals.

MACHINE LEARNING ENGINEER: *Byju's - Hyderabad, India* **Nov 2019 - Jun 2021**

- Developed Spark-based data pipelines for ingesting and cleaning large-scale student engagement data; reduced latency by 30%.
- Engineered deep learning pipelines and optimized ETL workflows for education data, improving accuracy by 15% on prediction models
- Deployed models in production using Flask APIs and container orchestration (Docker, Kubernetes), ensuring 95% deployment success.

PERSONAL PROJECTS

AUTONOMOUS GPT ENGINE - *Python, PyTorch, Transformers, BPE*

- Architected a 124M-param GPT model with transformer agents and BPE tokenizer for agentic decision-making in GenAI pipelines.
- Adapted transformer-based LLMs for classification and instruction tasks; achieved 90%+ accuracy, outperforming LLaMA 3.

RAG-DEEPSEEK: Conversational AI over PDFs - *LangChain, FAISS, ChromaDB, Deepseek, Streamlit*

- Implemented RAG-based document QA system over unstructured PDFs; reached 95% retrieval precision and 92% F1 on QA metrics
- Launched a scalable Streamlit interface with vector search over 1M+ PDFs, maintaining 99.9% uptime and sub-second latency.

FRAUD DETECTION & RISK ANALYSIS FOR PAYMENTS - *Python, SQL, Snowflake, Tableau, AWS Lambda*

- Designed anomaly detection system with Isolation Forest & One-Class SVM; reduced fraud by 40%, saving $500K per quarter.
- Streamlined fraud review workflows using Tableau and Lambda; reduced manual effort by 70% through live data visualizations.

EDUCATION

M.S. in Computer Science (Specialization: Artificial Intelligence) - University of South Dakota, USA **GPA: 4.0** **Aug 2023 - Dec 2024**

- **Courses:** ML, DL, Distributed Systems, Quantum Computing, Computer Vision, Applied Mathematics, HPC, Design Patterns, Big Data
- **Leadership:** President, Applied Artificial Intelligence (𝟚𝔸𝕀) Club, Feb 2025

B.S. in Computer Science - Lovely Professional University, India **GPA: 3.6** **Jul 2015 - May 2019**

- **Courses**: Programming, OS, Mathematics, SQL, Statistics, Linux, DSA, Version Control, CI/CD, Cloud Computing, Data Analytics