Chetan Valluru

valluruchetanreddy@gmail.com · 7325327879 · Portfolio · GitHub · LinkedIn

Summary

AI Engineer with 2+ years hands on experience in production-scale AI. Specialized in LLMs, generative AI, and MLOps. Improved model performance, cut inference costs. Deep knowledge of transformers, fine-tuning, and responsible deployment.

Professional Experience

AI Engineer August 2025 - Present

JerseySTEM, NJ

- Built a natural language to SQL agentic platform for self-serve data access; reduced ad-hoc data tickets by 30%, accelerated time-to-insight by at least 80%, and made insights accessible to non-technical teams.
- Implemented schema and data aware RAG to reliably handle complex multi-table functions, and ambiguity.
- Shipped AI agents with conversational flows, and context control for accurate knowledge retrieval and guided workflows.
- Led integrations across Google Workspace, Slack, Salesforce, and Jira; added testing, monitoring, and docs to drive adoption.

Machine Learning & Statistics Grader September 2023 – May 2024

Rutgers University, NJ

- Evaluated 60 weekly student assignments on statistical inference, hypothesis testing, and machine learning, delivering actionable feedback to improve project reproducibility and rigor.

AI/ML Research Assistant August 2021 – May 2023

VNR VJIET, India

- Developed multilingual NLP system fine-tuning mBERT and IndicBERT transformers, improving information accessibility for 50,000+ farmers and reducing language barriers by 40%.
- Architected data collection infrastructure using Python, Scrapy, and BeautifulSoup, building domain-specific corpus of 2M+ documents and accelerating regional language research by 60%

Technical Skills

Artificial Intelligence: Transformers, LangChain, Pinecone, Prompt Engineering, RAG, LoRA, Knowledge Distillation, Quantization-Aware Training, Generative AI, PydanticAI, n8n, Weights & Biases

Machine Learning & Deep Learning: Classification, Regression, Clustering, Ensemble Methods, Neural Networks, BERT, LSTM, GANs, NLP, PyTorch, TensorFlow, Scikit-learn, Reinforcement Learning

Data Engineering & Cloud: Python, SQL, Data Warehousing, Agile Methodologies, Git, Data Pipelines, GCP Vertex AI

Education

Rutgers University September 2023 – May 2025

Master of Science in Data Science, New Brunswick, NJ GPA: 3.65/4.0

VNR VJIET August 2019 – May 2023

Bachelor's degree in Electronics and Communications with Machine Learning and AI, India

Projects

Taylor: Your AI Career Stylist — Full Stack LLM Application Check it out!

- Engineered web-based LLM application generating personalized resumes and cover letters, serving 250+ active users and reducing application processing time by 50%

GemmA.I: Private On-Device Vision Assistant — iOS, Local AI, Accessibility Watch the Demo!

- Built an iOS app for blind/low-vision users with real-time scene understanding, obstacle detection and spoken assistance, featuring 100% on-device processing for maximum privacy and mobile-optimized performance.

BetterLlama: Smaller and better LLM — LLM Optimization, Security, Ethical AI

- Fine-tuned LLaMA 3.2 with advanced quantization techniques (8-bit, LoRA), achieving 1.8x faster inference while maintaining 99.9% accuracy. Integrated chain-of-thought prompting and safeguards to ensure secure, responsible AI usage.

ChessGPT: Transformers Based Chess Engine — NLP, Large Language Model

- Built a transformer model from scratch for move-sequence generation, producing 30 legal moves per prompt with a small model footprint. Demonstrated data tuning techniques outperform naive fine-tuning and RAG approaches for language modeling.

NJ Transit Smart Journey Assistant — Machine Learning, RAG, AI

- Developed rail-delay prediction pipeline with dataset of NJ Transit delay data to reduce operational delays by 10%. And integrated RAG chatbot using NJ Transit FAQ's data

Publications & Achievements

- *V. Chetan Reddy et al., "Multi-Classification of Respiratory Diseases using Deep Learning," 2023 International Conference on Sustainable Computing and Smart Systems, DOI: 10.1109/ICSCSS57650.2023.10169597* [AI/ML Domain]