Ajay Therala

+1 (623) 286-1552| atherala@asu.edu | ajay-therala7 | Ajaytherala

Education

Master of Science, Computer Science — **August 2023 - May 2025**
Arizona State University, Tempe, AZ, United States. — **GPA: 3.93/4.0**
Course Work: Topics in NLP, Data Mining, Data Visualization, Statistical ML, Data Processing at Scale, Knowledge Representation & Reasoning.

Bachelor of Technology, Information Technology — **August 2017 - July 2021**
Jawaharlal Nehru Technological University, Hyderabad, TS, India — **GPA: 9.21/10.0**
Course Work: DSA Using C, Java & Python, Operating Systems, Object Oriented Programming, Linear Algebra, Probability & Statistics, DBMS

Technical Skills

Programming Languages: Python, SQL, JavaScript, C.
AI & Machine Learning: Generative AI, LLMs (LangChain, LangGraph, Hugging Face), **TensorFlow**, **PyTorch**, scikit-learn, Deep Learning, Pandas, NumPy, NLP.
Data Engineering & Big Data: PySpark, **Spark SQL**, Hadoop.
Frameworks & Databases: Django, Streamlit, Tableau, PostgreSQL, MongoDB, DynamoDB, Vector Databases (OpenSearch).
Cloud & DevOps: AWS (S3, Lambda, SageMaker, Bedrock, API Gateway, ECR), CI/CD Pipelines, Terraform, Docker, Jenkins, GitHub.

Certifications

- **Microsoft Certified: Azure Data Scientist Associate** *Verify Here*.
- **Oracle Certified : Oracle Cloud Infrastructure 2024 Generative AI Certified Professional** *Verify Here*.

Publications

[1] **Evaluating Multimodal Large Language Models across Distribution Shifts and Augmentations, Publisher : IEEE** *Paper Link*.
[2] **NetMD- Network Traffic Analysis and Malware Detection, Publisher: IEEE** *Paper Link*.

Professional Experience

AI Data Engineer, Dexian LLC — **May 2025 - Present**
Arizona State University (Client) — **Tempe, United States.**
- **Architected** conversation truncation logic to intelligently manage token context windows, eliminating context overflow errors and ensuring robust LLM performance in production workloads.
- **Spearheaded** the end-to-end design and development of a universal URL processing **API** with **integrated caching**, expanding platform capabilities from handling specific domain to any web URL.
- **Deployed** a scalable ingestion API to onboard **25,000+ Canvas courses** into OpenSearch for RAG pipelines, establishing **role-based access control (RBAC) by provisioning professors as owners and students as viewers**.

AI Full Stack Developer, Arizona State University — **August 2024 - May 2025**
AI Acceleration Team — **Tempe, United States.**
- **Optimized Data Ingestion Pipelines** - Spearheaded a **95% reduction** in file chunking times for large datasets (from 1500 s to 4.87s) by integrating an **advanced semantic chunking**, accelerating downstream embedding and retrieval efficiency.
- **Streamlined** end-to-end deployment via a**CI/CD pipeline** - iteratively tested in Sage Maker, managed GitHub version control, dockerized applications, published images to AWS ECR, & deployed to Lambda for scalable serverless execution.
- **Engineered** a **RAG pipeline** by integrating **Google Drive APIs** for document extracting, chunking, embedding, and **Open Search indexing**, enabling fast, precise Retrieval-Augmented Generation.

Systems Engineer (ML Developer), Tata Consultancy Services Limited — **August 2021 - August 2023**
TechBU, Digital Research & Innovation — **Hyderabad, India**
- Engineered core components for a domain-specific search engine (Cognitive Product Support), **enhanced search accuracy by 30%**.
- Crafted Data Lens, a component for training custom NER models, achieving an **impressive 80% - 90% accuracy**.
- Demonstrated expertise in Generative AI, **Prompt Engineering** by developing advanced GPT-powered bots handling over **5,000+ interactions** daily. Delivered **impactful client demos**, earning high praise from esteemed clientele.
- **Led** the development of an automated document digitization by developing **Textract-based processing solutions** and prototyping a system for handwritten form extraction, reducing manual effort and turnaround time.

ML Research Project Intern — **Jan 2021 - Aug 2021**
Tata Consultancy Services Limited — **Hyderabad, India**
- Enhanced intrusion detection performance by **6%** on NetML, CICIDS2017, and non-vpn2016 datasets using **data refinement techniques**, and **Bagging & Boosting** algorithms.
- **Achieved** a Top-5 global rank in the NetML Network Traffic Analytics Challenge 2020 and presented research findings at **ICAIIC 2022** to over 300+ peers and industry professionals.
- Orchestrated end-to-end machine learning workflow using Django, encompassing ingestion, EDA, model training, and evaluation showcasing strong ML Engineering & deployment practices.

Projects

AI Resume Analyzer.
- Developed an **AWS-powered AI Resume Analyzer** leveraging **Amazon Bedrock's LLM**, API Gateway, and Lambda to assess job fit, identify skill gaps, and provide **AI-driven** career insights through an interactive Streamlit UI.

Business & User Level Analysis of YELP Dataset.
- **Optimized** large-scale Yelp dataset processing with **PySpark** & **Spark SQL**, building scalable **ETL** pipelines, execuring distributed queries, and created interactive **Tableau** dashboards to uncover key business and user trends.

Data Analyst (Volunteer Research Assistant)
- **Analyzed** GB's of EEG signals, physiological and sensor data from 32 participants under 6 conditions, using **statistical modeling** & **visualization**, evaluating **cognitive performance under varying insulation materials & temperatures**.

Evaluating Multimodal Large Language Models across Distribution Shifts and Augmentations.
- **Benchmarked** state-of-the-art MLLMs (InstructBLIP, LLaMA) across VQAv2 and CLEVR datasets under multimodal distribution shifts, revealing **key robustness gaps** in synthetic and complex reasoning settings.
- Generated **75K+ QA pairs** and conducted fine-grained analysis identifying performance degradation due to visual perturbations and logical connectives, especially in color and count-based tasks.