Sandeep Kalari

skala003@odu.edu — +1-571-752-2805 — Norfolk, VA
LinkedIn — GitHub — Portfolio

Summary of Qualifications

- Ph.D. candidate (ABD) with 5+ years of research experience in Trustworthy AI, Multimodal LLMs, and Explainable AI systems.
- Proven track record of developing and deploying end-to-end AI solutions achieving up to 88% accuracy in production environments.
- Expertise in Retrieval-Augmented Generation (RAG), Vision-Language Models (VLMs), and Parameter Efficient Fine-Tuning (PEFT) techniques including LoRA and QLoRA.
- Published 7+ peer-reviewed papers at top-tier IEEE/ACM conferences on AI applications in e-commerce, blockchain, and aerospace.
- Leadership experience as President of Computer Science Graduate Society, managing cross-functional teams of 10+ members.

Education

Old Dominion University, Norfolk, VA
Ph.D. in Computer Science (All But Dissertation) — Expected Dec 2026
Research Focus: Trustworthy AI, Multimodal Large Language Models, Explainable AI
Master of Science in Computer Science (GPA: 3.97/4.0) — Dec 2023

CBIT, Osmania University, Hyderabad, India
Master of Computer Applications (MCA) — Aug 2021

Technical Skills

Programming Languages: Python, Java, R, JavaScript, SQL, C++, CUDA, HTML/CSS
AI/ML/DL Frameworks: PyTorch, TensorFlow, Keras, Scikit-learn, Transformers, LangChain, OpenCV
Large Language Models: GPT-4, Claude, LLaMA 3.2, Qwen 2v 7B, Phi-3.5, RAG pipelines
Cloud & DevOps: AWS (Bedrock, EC2, S3, OpenSearch, QuickSight), Docker, Kubernetes, CI/CD
Data Engineering: Pandas, NumPy, Spark, Hadoop, Kafka, Vector Databases, ETL pipelines
Specialized Areas: Vision-Language Models, Prompt Engineering (CoT, Few-Shot), PEFT (LoRA, QLoRA), Multimodal AI

Professional Experience

AI Solutions Engineer Intern — May 2025 – Jul 2025
TechAbstractor LLC, Arlington, TX

- Engineered and deployed full-stack AI chatbot serving as virtual CFO, **reducing manual financial evaluation time by 75%** for accounting firms.
- Architected scalable AWS infrastructure (EC2, S3, OpenSearch, Bedrock) with load balancing, achieving **99.9% uptime**.
- Developed sophisticated RAG pipeline using Claude LLM, automating financial data categorization with **92% accuracy**.
- Implemented advanced prompting techniques (Chain-of-Thought, Few-Shot), improving response accuracy by **35%**.
- Built React/Node.js application integrating third-party APIs and AWS QuickSight for real-time data visualization.

Graduate Research Assistant — Feb 2022 – Present
Old Dominion University, Norfolk, VA

- Developed novel Vision-Language Model achieving **88% accuracy** in detecting e-commerce review discrepancies, outperforming LLaMA 3.2 and Phi-3.5 baselines.
- Fine-tuned Qwen 2v 7B using PEFT techniques (LoRA, QLoRA), **reducing training time by 60%** while maintaining model performance.
- Built multimodal AI framework processing video, audio, text, and sensor data for aircraft maintenance with **86.44% classification accuracy**.
- Trained and deployed 10+ ML/DL/LLM models, including custom chatbots using RAG architecture for domain-specific applications.
- Designed blockchain-integrated LLM framework ensuring data provenance with **86.25% privacy preservation rate**.
- Published 5+ papers at IEEE/ACM conferences, contributing to state-of-the-art in trustworthy AI research.

ML Engineering Intern Aug 2021 – Dec 2021
Infinity Hydroponic Green Farms, Hyderabad, India

- Built end-to-end IoT automation prototype integrating Raspberry Pi sensors with lightweight ML models (YOLO, BERT) for real-time monitoring.
- Achieved **95% accuracy** in plant health detection using computer vision on water/weather images, enabling predictive notifications.
- Developed autonomous notification system for nutrient levels, temperature, and plant health, **eliminating 100% manual monitoring**.
- Deployed cross-platform alert system accessible on any device, reducing response time to critical events by **80%**.

Research Assistant June 2020 – Aug 2021
Chaitanya Bharathi Institute of Technology, India

- Developed IoT-based precision farming system with ML integration, **increasing crop yield by 30%**.
- Published 2 research papers on automated agriculture systems leveraging IoT and AI technologies.

Selected Publications & Projects

ReViewQwen: Explainable VLM for E-Commerce Review Analysis IEEE CBMI 2025

- Achieved **88% accuracy** detecting seller-buyer discrepancies using novel multimodal framework.
- Outperformed baseline models (LLaMA 3.2, Phi-3.5) by 12% through advanced prompt engineering.

BlockQwen: Blockchain-Powered LLM for Trustworthy AI IEEE BCCA 2025

- Integrated Qwen LLM with smart contracts achieving **86.25% privacy preservation** in simulations.
- Reduced data tampering risks by 90% through blockchain-based audit trails.

AeroQwen: Multimodal Framework for Aircraft Maintenance IEEE Aerospace 2026

- Integrated video, audio, text, and sensor data streams for comprehensive component health assessment with **86.44% accuracy**.
- Processed 100GB+ synchronized multimodal data using optimized multi-GPU architecture and Chain-of-Thought reasoning.

MerkleChain: Privacy-Enhanced Blockchain Framework In Submission

- Developed and implemented prototype achieving **3x faster transaction processing** through parallel Merkle tree operations.
- Integrated chameleon hashing with ECC enabling GDPR-compliant data redaction while maintaining **100% tamper-evidence**.
- Reduced miner storage requirements by **70%** through selective data subscription model.

Additional Publications: 3+ papers on web accessibility and digital libraries at ACM/IEEE venues.

Leadership & Awards

- **President**, Computer Science Graduate Society, ODU (2024-Present) - Led 10+ member team
- **Graduate Teaching Assistant**, ODU - Mentored 100+ students in Data Structures and Algorithms
- **Research Excellence Award**, ODU Computer Science Department (2024)