Dr. Phani Siginamsetty

siginamsettyphani@gmail.com | +91-8125636250 | LinkedIn | GitHub | Google Scholar | Portfolio

PROFESSIONAL SUMMARY

Associate Data Scientist and PhD Researcher specializing in Generative AI, Machine Learning, and autonomous Multi-Agent Systems. Proven track record of bridging academic research with high-impact enterprise solutions, focusing on orchestration frameworks where autonomous agents execute complex workflows. Deep expertise in end-to-end LLM lifecycle management, including RAG pipelines, fine-tuning (PEFT/LoRA), and quantized edge deployment.

TECHNICAL SKILLS

- **GenAI & Agentic Frameworks:** Multi-Agent Orchestration (CrewAI, AutoGen, LangGraph, Agno), PydanticAI, Haystack, Tool-Function Calling, Semantic Routing, HITL, Evaluation (Ragas, TruLens).
- **LLMs & Model Engineering:** AWS Bedrock, Llama 3.2, GPT-4o, Fine-Tuning (PEFT, QLoRA), Unsloth, Hugging Face TRL (DPO/RLHF), DeepSpeed, FSDP, Model Quantization (GGUF, AWQ, GPTQ).
- **Data Science & Advanced ML:** PyTorch, Scikit-Learn, Reinforcement Learning (PPO, DQN), Time-Series, XGBoost, Random Forest, Siamese Networks, AutoML, Quantum-Motivated Algorithms.
- **Vision & Multimodal AI:** Multimodal RAG, OpenCV, AWS Textract (OCR), Document Intelligence, OpenAI Whisper (ASR), ElevenLabs (TTS), CLIP, Image & Audio Processing.
- **MLOps, Cloud & Backend:** Python (Expert), AWS (SageMaker, Bedrock, Lambda, EC2), Docker, Kubernetes, CI/CD, Git, FastAPI, Flask, RESTful APIs, JWT/RBAC, Optuna.
- **Data Infrastructure:** Vector DBs (Pinecone, Milvus, Weaviate, Chroma, FAISS), PostgreSQL, MongoDB, Spark, Pandas, NumPy, Parquet, Data Encryption.

WORK EXPERIENCE

Hexaware Technologies — Chennai, India
Associate Data Scientist — *March 2025 – Present*

- **Autonomous Fraud Detection:** Spearheading real-time fraud detection using stateful multi-agent systems via the **Agno** framework with advanced tool-calling for complex transaction analysis.
- **Advanced RAG Pipelines:** Engineering a multi-agent RAG pipeline for "Smart Tutor" using vector databases and semantic routing to deliver personalized content while minimizing hallucinations.
- **Enterprise Automation:** Designing agentic workflows with LangChain and LangGraph to automate reporting with Human-in-the-Loop (HITL) mechanisms, reducing operational overhead.

Volvo Group — Bangalore, India
Research Assistant — *June 2024 – March 2025*

- **Edge GenAI & Quantization:** Researched lightweight LLMs for on-device inference using GGUF/AWQ quantization to reduce memory footprint and latency on vehicular hardware.
- **Computer Vision Diagnostics:** Deployed an optimized CNN pipeline for real-time component recognition within the Vehicle Configuration Manager (VCM) to automate visual inspections.

SRM University AP — Amaravati, India
Data Science Researcher (PhD Scholar) — *Sep 2021 – July 2024*

- **Healthcare AI Consultant:** Architected a secure RAG chatbot for SRM Global Hospital to retrieve medical protocols while ensuring strict data privacy and proprietary data embedding.
- **Audio Intelligence:** Engineered a MoM automation API using **FastAPI**, STT, and speaker diarization to autonomously extract abstractive summaries and action items from recordings.
- **Multilingual NLP:** Developed MATSFT and MMSFT frameworks by fine-tuning mT5 for low-resource Indian languages, resulting in multiple high-impact journal publications.
- **Quantum AI & IP:** Architected quantum-motivated summarization processors for data compression, leading to multiple Indian Patents, including **1 Granted Patent**.

Tychee Innovations — Andhra Pradesh, India
Trainee Engineer (ML Research) — *Aug 2020 – Jul 2021*

- **Industrial Safety Vision:** Deployed real-time object detection to monitor hazardous machinery, triggering emergency stops via spatial tracking of hand proximity to danger zones.
- **Predictive Analytics:** Developed ML models to forecast patient outcomes and translate clinical data into actionable insights for data-driven healthcare decisions.

Dhanekula Engineering College — Andhra Pradesh, India
Assistant Professor — *Oct 2020 – Aug 2021*

- **Software Mentorship:** Instructed Data Structures, Algorithms, and Python, mentoring students in software engineering best practices and technical problem-solving.

KEY PROJECTS

SmartTutor: Multimodal RAG Learning Platform — *FastAPI, AWS Bedrock, Pinecone, AWS Polly*

- **Multimodal Ingestion Engine:** Architected an advanced RAG pipeline parsing unstructured text, complex PDFs, and diagrams using AWS Bedrock (Nova Pro & Claude 3.5) for curriculum-aligned content generation.

- **Context-Aware Retrieval:** Orchestrated a scalable vector search architecture using Pinecone with semantic routing and hybrid search (dense + sparse) to drastically reduce retrieval latency.
- **Interactive Agentic Loop:** Engineered an autonomous, stateful assessment agent that adapts to student performance with real-time TTS auditory feedback via AWS Polly.

Enterprise Fraud Prevention System (Citi Bank) *Python, Scikit-Learn, Multi-Agent Orchestration, AWS*

- **Hybrid Risk Intelligence:** Spearheaded a dual-layered risk engine fusing statistical anomaly detection (XGBoost) with GenAI-driven forensics, reducing investigation time and false positives.
- **Autonomous Rule Discovery:** Innovated a multi-agent workflow for live transaction monitoring, utilizing tool-calling to detect zero-day fraud patterns with Human-in-the-Loop oversight.
- **Secure Data Parsing Agent:** Designed "Argus," an AI Data Analyst using `msoffcrypto` to securely decrypt and parse sensitive financial datasets locally for evidence-based risk verdicts.

Automated Bank Cheque Verification System *Computer Vision, PyTorch, AWS Textract, OpenCV*

- **Forensic Digitization Pipeline:** Constructed an end-to-end vision pipeline using AWS Textract and OpenCV for layout analysis and digitization of MICR codes and payee details with high OCR accuracy.
- **Signature Verification:** Engineered a PyTorch-based Siamese Neural Network for one-shot learning, utilizing contrastive loss and feature embeddings to detect sophisticated forged signatures.
- **Cross-Modal Logic:** Programmed NLP algorithms to cross-verify extracted semantic data (e.g., matching numeric amounts against written text) to flag discrepancies for manual review.

Personalized Medical AI Assistant *Llama 3, Agno, LangGraph, FastAPI, MongoDB*

- **Clinical Guardrails & RAG:** Engineered a domain-specific agent using Llama 3 and Vector DBs, implementing query expansion and re-ranking to ground answers exclusively in verified medical literature.
- **Stateful Memory Architecture:** Developed a persistent context-retention engine using LangGraph and MongoDB to map longitudinal symptoms and medical history for personalized health insights.

EDUCATION

Ph.D. in Computer Science & Engineering 2021 – 2025
SRM University AP *CGPA: 9.25/10.0*

M.Tech in Computer Science & Engineering 2018 – 2020
KL University *CGPA: 8.5/10.0*

B.Tech in Computer Science & Engineering 2014 – 2018
JNTUK *81.0%*

Intermediate (MPC) 2012 – 2014
Board of Intermediate Education *91.70%*

SSC (10th Standard) 2011 – 2012
Board of Secondary Education *GPA: 9.3/10.0*

PATENTS

- System and a method for automated exam evaluation and personalized learning feedback
  Indian Patent No: 202541018210 | *Indian Patent Journal*, 2025
- A System and a Method for Managing API Calls in A Large Language Model
  Indian Patent No: 202441096836 | *Indian Patent Journal*, 2024
- A System and a Method for Healthcare Data Processing and Decision Support
  Indian Patent No: 202441076761 | *Indian Patent Journal*, 2024
- System and method for multilingual fake news detection in multimodal information
  Indian Patent No: 202441030030 | *Indian Patent Journal*, 2024
- A Healthcare Summarization System and A Method Thereof
  Indian Patent No: 202441005845 | *Indian Patent Journal*, 2024
- A System and a Method for Personalized E-Content Generation Based on Student Performance
  Indian Patent No: 202441003347 | *Indian Patent Journal*, 2024
- System and method for deriving multilingual meeting minutes
  Indian Patent No: 202441001022 | **Grant No: 581292** | *Indian Patent Journal*, 2024
- System and method for multimodal multilingual input summarization using quantum motivated processors
  Indian Patent No: 202341005519 | **Grant No: 66614** | *Indian Patent Journal*, 2023
- A System and A Method for Generating Trading Coupons
  Indian Patent No: 202341007665 | *Indian Patent Journal*, 2023
- A System and A Method for Prediction of The Strength of Concrete
  Indian Patent No: 202341007257 | **Grant No: 582851** | *Indian Patent Journal*, 2023
- A System and Method for Performing Multilingual Multimodal Summarization
  Indian Patent No: 202241073648 | *Indian Patent Journal*, 2022

PUBLICATIONS

- MATSFT: User query-based multilingual abstractive text summarization for low resource Indian languages by fine-tuning mT5
  Phani, S., et al. | *Alexandria Engineering Journal*, Elsevier, 2025. DOI: 10.1016/j.aej.2025.04.031

- Improving Preliminary Clinical Diagnosis Accuracy through Knowledge Filtering Techniques in Consultation Dialogues
  Abdul, A., **Phani, S.**, et al. | *Computer Methods and Programs in Biomedicine*, Elsevier, 2024. DOI: 10.1016/j.cmpb.2024.108051

- MMSFT: Multilingual Multimodal Summarization by Fine-tuning Transformers
  Phani, S., et al. | *IEEE Access*, 2024. DOI: 10.1109/ACCESS.2024.3454382

- MMSML: Multilingual Multimodal Summarization for Multimodal Input
  Phani, S., et al. | *Intl. Conference on Data Science and Applications*, Springer, 2024. DOI: 10.1007/978-981-96-2724-0_5

- Recognition for Attendance System Using Reinforcement Learning
  Phani, S., et al. | *FICTA*, Springer, 2023. DOI: 10.1007/978-981-99-6702-5_15

- Abstractive Text Summarization with Fine-Tuned Transformer
  Phani, S., et al. | *MAI 2022*, Springer, 2023. DOI: 10.1007/978-981-99-0189-0_46

- Machine Learning Classifiers and Along with TPOT Classifier (AutoML) to Predict the Readmission Patterns of Diabetic Patients
  Phani, S., et al. | *IJRTE*, 2020. DOI: 10.35940/ijrte.f7415.059120