

Phani Srikanth

Applied AI, Evals and Agentic Systems

Hyderabad, India | phanisrikanth33@gmail.com | phanisrikanth.com | linkedin.com/in/phanisrikanth | github.com/binga

SUMMARY

Applied AI and research leader with 12+ years building the context, data, and evaluation systems that make AI useful in production. Leads NetApp's AI Data Engine work across semantic data intelligence, multimodal understanding, synthetic data, LLM-assisted labeling, retrieval, governance, and inference optimization, while building and leading applied-science teams.

EXPERIENCE

Principal Scientist & Director, Applied AI | India AI/ML Site Lead *NetApp*

May 2023 – Present
Hyderabad, India - Remote

- Lead AI/ML research for NetApp's flagship AI Data Engine, which aims to build a "Context Layer for AI" by transforming storage into an intelligent semantic file system that makes enterprise data discoverable, governed, and agent-ready across on-premises and cloud environments.
- Built storage-scale classification spanning 200 entity types and processing 200K files/day on a single L4 GPU, enabling low-cost semantic search, retrieval, sensitive-data detection, and governed access across structured and unstructured data.
- Improved model accuracy from 61% to 85% through synthetic-data generation and LLM-generated ground truth, supported by rigorous evaluations and inference optimizations for production deployment.
- Scaled NetApp's India AI/ML team from 1 to 10 applied scientists, resulting in 10+ patent filings and A* conference publications; presented applied AI work at NetApp INSIGHT 2024 and 2025.

Principal Applied Scientist & Manager *Microsoft*

Mar 2019 – May 2023
Hyderabad, India & Redmond, WA, USA

- Shipped machine-learning models to production and led teams delivering applied-ML systems within Microsoft Security Research.
- Worked across Microsoft Defender for Endpoint, Web, and Email teams, spanning principal individual-contributor and people-manager roles.
- Built experience at the intersection of endpoint, web, and email security, production ML, and applied research across India and the United States.

Machine Learning Consultant *Atrium (YC W18)*

Apr 2018 – Aug 2018
Remote

- Delivered a working legal language model by fine-tuning an AWD-LSTM architecture on EDGAR filings and implementing end-to-end inference for domain-specific text generation.
- Enabled scalable legal-clause retrieval with document and query embeddings using the legal language model and ANN search for fast retrieval.

Earlier Data Science Roles

Sep 2014 – Jan 2019

- Reliance Jio:** Built its first 100+ node on-premises data platform and grew the data science and engineering function from 1 to 10+ as Jio launched and scaled to 100M DAU in months.
- Housing.com:** Built ML models to identify stale real-estate listings.
- iCube:** Shipped AutoML software that accelerated cross-industry ML deployments.

SELECTED RESEARCH & OPEN SOURCE

ICDAR 2026 DocVQA ([code](#), [report](#)) 2026
Raised document-VQA validation accuracy from 3.75% to 46.25% by evolving zero-shot VLM inference into an agentic REPL with code execution, OCR, visual navigation, search, and answer repair.

ICDAR 2026 Writer Identification ([code](#)) 2026
Improved private accuracy from 0.386 to 0.481 (+9.56 points) and placed 6th of 113 teams using leave-writer-out evaluation, conservative pseudo-labeling, three-seed ensembling, and calibrated open-set rejection.

Predict the LLM ([code](#)) 2023
Placed 2nd of 106 teams by combining GroupKFold evaluation with multi-backbone decoder ensembles, improving leaderboard log loss from 0.623 to approximately 0.500.

GPU MODE Kernel Competition ([code](#)) - Optimized GPU kernels using autoresearch techniques. 2026

TECHNICAL FOCUS

Semantic retrieval and ANN | Multimodal Document AI | Synthetic data and LLM-assisted labeling | Evaluation harnesses | Data classification and governance | PyTorch, Modal

EDUCATION

Deep Learning, fast.ai, University of San Francisco, USA 2018
Bachelor's in Electrical Engineering, National Institute of Technology Warangal, India 2012

PUBLICATIONS AND PATENTS

Open Your Eyes: Benchmarking Fabricated Realities and Weaponized Ethics in VLMs... ECCV, 2026
Data Preparation Engine(s) for Secure and Compliant Data Collections..., US-20260080099-A1 2025
Additional patent filings across cybersecurity and Document AI 2024–2025