Kamesh R

+917299959898 | kameshrajeshkanna@outlook.com | LinkedIn

Education

University College London (UCL) — London, UK
Master of Science in Machine Learning — *Aug 2026 – Aug 2027*

Sathyabama Institute of Science and Technology — Chennai, India
Bachelor of Engineering in Computer Science (AI and Machine Learning) — *Sep 2022 – May 2026*

Publications

Kamesh R. *Probing Reasoning Flaws and Safety Hierarchies with Chain-of-Thought Difference Amplification. NeurIPS 2025, LLM-Evals Workshop.*

Kamesh R. *Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures.*

Kamesh R. et al. *Indian Grammatical Tradition-Inspired Universal Semantic Representation Bank (USR Bank 1.0). AACL-IJCNLP 2025, BHASHA Workshop.*

Jerome Francis, **Kamesh R.**, **Serena Pei.** *PerplexMATH: Steering LLMs Toward Mathematical Reasoning. ICML 2025, NewInML Workshop (Poster Presentation).*

Experience

Machine Learning Intern — Feb 2026 – Present
Glance — *Bengaluru, India*

- Built and optimized catalog retrieval pipelines for the Glance app, improving content discovery across millions of items.
- Researched and reimplemented SOTA embedding techniques from recent papers to develop internal catalog-focused embeddings tailored to Glance's content domain.
- Achieved a 2x improvement in retrieval metrics by reimplementing SOTA techniques, significantly enhancing catalog search relevance and ranking quality.

Machine Learning Contributor (Contract) — Feb 2026 – Present
Shipd by Datacurve (YC W24) — *Remote*

- Participated as a key contributor in high-stakes Machine Learning challenges for a YC-backed data platform.
- Awarded monetary prizes for top-tier performance in predictive modeling and algorithmic problem-solving tasks.
- Developed and optimized ML models under strict competitive constraints, delivering results in a fast-paced environment.

Research Fellow — Jan 2026 – March 2026
AI Safety Camp (AISC) — *Remote*

- Working with **Ihor Kendiukhov** to evaluate the scalability and security guarantees of novel control protocol classes for advanced AI systems.
- Extending Greenblatt et al.'s framework by designing hierarchical and parallel control topologies to optimize safety-usefulness trade-offs.
- Conducting scaling experiments to verify the generalization of control guarantees across diverse model capabilities.

Independent Interpretability Researcher — Nov 2025 – Feb 2026
Mentored by David Africa (UK AI Security Institute) — *Remote*

- Conducting independent research on mechanistic interpretability, with direct technical guidance and mentorship from a Research Scientist at the UK AI Safety Institute.
- Developing a framework to permanently embed steering vectors into model weights, enabling persistent safety behaviors without inference-time compute overhead.

Research Intern — May 2025 – Oct 2025
International Institute of Information Technology, Hyderabad — *Hyderabad, India*

- Researched Universal Semantic Representations and developed techniques for generating coherent natural language sentences from abstract syntactic-semantic structures.

- Worked on Controlled Image-to-Text Generation systems for scientific images to ensure accurate, context-aware, and domain-specific textual descriptions.
- Conducted workshops on prompt engineering for linguistics researchers, introducing strategies to effectively use large language models for linguistic data processing and analysis.

Machine Learning Intern Dec 2024 – Jan 2025

Co-build.tech *Remote*

- Enhanced query performance by 25% by optimizing document embedding strategies for semantic similarity retrieval.
- Automated mapping of legal clause dependencies, improving analysis efficiency by 30%.
- Conducted vulnerability analysis to identify and mitigate critical risks in legal document workflows.

Grants

Lambda Labs Research Grant — Awarded $1,000 in initial GPU credits (scalable up to $5,000) for conducting research on the interpretability and routing dynamics of Mixture-of-Experts (MoE) models during inference. Work involves mathematical analysis of expert load balancing, probing memorization vs. generalization through expert usage patterns, and inference-only diagnostic tools.

Service

Reviewer - NeurIPS'25 AI-MATH Workshop

Reviewer - AAAI'26 Workshop on Shaping Responsible Synthetic Data in the Era of Foundation Models

Leadership & Mentoring

Peer Education Initiative Fall 2023 – Winter 2024

Technology Hub, Sathyabama Institute of Science and Technology

- Organized and delivered supplementary ML and Deep Learning workshop series for 26 students, covering supervised/unsupervised learning, CNNs, RNNs, and transfer learning.
- Designed curriculum materials and provided hands-on guidance for implementing real-world applications including object detection, NLP tasks, and GANs.
- Delivered guest seminars on neural networks for first-year students, mentoring peers through their initial machine learning projects.

Projects

Black Scholes Modelling with Kolmogorov Arnold Networks

- Currently working to improve the 72% accuracy achieved on the Synthetic European dataset by refining Kolmogorov Arnold Networks for better modeling of financial derivatives.
- Tracked and optimized model performance, with final metrics showing a train loss of 2.84 and a test loss of 4.57.
- Managed regularization with a final value of ± 2.05, improving model stability and generalization.

Internalizing Safety via Weight-Level Activation Steering

- Developed a novel model editing technique to permanently "bake" transient safety steering vectors into Transformer model weights, transforming temporary activation interventions into persistent architectural alignment without inference-time overhead.
- Achieved robust safety behavior comparable to standard runtime steering while preserving general coherence, effectively internalizing the safety concept directly into the target architecture's residual stream projections.

Talks & Seminars

Presentation – Controlled Image Generation 2025

Language Technologies Research Center (LTRC), IIIT Hyderabad *Hyderabad, India*

Talk & Workshop – Prompt Engineering 2025

Language Technologies Research Center (LTRC), IIIT Hyderabad *Hyderabad, India*

Lecture – Safety and Alignment of LLMs 2025
Sathyabama Institute of Science and Technology *Chennai, India*

Panelist – Impact and Application of GenAI 2024
Google Developer Summit *Chennai, India*

Speaker – Introduction to Machine Learning and Scopes 2024
Sathyabama Institute of Science and Technology *Chennai, India*

Technical Skills

Languages & Tools: Python, C++, SQL, JAX, Linux, WandB,Modal
ML Frameworks: PyTorch, JAX, Transformers (HF), DeepSpeed, vLLM, TRL, Unsloth
Core Libraries: NumPy, SciPy, Pandas, Scikit-learn, Matplotlib, Seaborn, Plotly, Datasets (HF)