Amaljith Kuttamath

Virginia, USA | (703) 969-6510
kuttamath.amaljith@gmail.com | linkedin.com/in/amaljithk | github.com/amaljithkuttamath | amaljithkuttamath.github.io

Professional Summary

AI Engineer with 7+ years of experience building production AI systems and researching LLM trust and safety. Currently focused on mechanistic interpretability, hallucination detection, and failure mode analysis at the architectural level. Proven track record deploying RAG systems, LLM evaluation frameworks, and knowledge graphs in healthcare. Building tooling for LLM trustworthiness profiling and model improvement.

Technical Expertise

AI Safety & Evaluation: LLM Trust Evaluation, Hallucination Detection, Mechanistic Interpretability, Model Calibration, Safety Benchmarking
LLM Engineering: RAG Systems, LLM Fine-tuning (LoRA, GRPO), Inference Optimization (vLLM, KV Cache), Agentic AI, Knowledge Graphs
ML Stack: Python, PyTorch, Hugging Face Transformers, lm-eval-harness, TrustLLM, MLflow, Vertex AI
Infrastructure: GCP, Azure, Databricks, Docker, Kubernetes, CI/CD, Elasticsearch, Neo4j
Full-Stack: Next.js, React, FastAPI, Flask, REST APIs, TypeScript

Professional Experience

Sorcero **Washington, DC**
AI Engineer *May 2024 – Present*

- Built AI evaluation framework combining LLM-as-judge methodology with human review, reducing model hallucination rates by 35% and establishing systematic trust measurement for production models
- Architected production medical RAG system on Vertex AI with hallucination-aware retrieval, achieving 30% improvement in retrieval fidelity across 10,000+ daily clinical queries
- Designed model monitoring and drift detection pipelines, identifying degradation patterns in LLM outputs before they reached end users
- Established AI governance frameworks with product, engineering, and compliance teams, ensuring model safety in regulated healthcare environments

Senior NLP Research Fellow *Jan 2024 – May 2024*

- Conducted feasibility studies for RAG implementations, evaluating retrieval accuracy and hallucination rates across medical document corpora
- Built proof-of-concept knowledge graph systems for medical entity extraction, testing model reliability on domain-specific content

INECTA (Microsoft Gold Partner) **Remote, USA**
Technical Consultant *May 2021 – Aug 2022*

- Engineered RAG-powered business intelligence platform enabling natural language querying of ERP data, serving 50+ enterprise clients
- Architected 8+ ETL pipelines using Azure Data Factory and Databricks, processing 2TB+ daily with 99.9% uptime
- Optimized data processing workflows resulting in 60% reduction in query response times and $200K annual cost savings

RadianArc Technologies **Mumbai, India**
Co-Founder *May 2019 – Apr 2021*

- Co-founded industrial AI company building digital twins for predictive maintenance on large-scale machinery using IoT sensor data
- Developed predictive maintenance platform (A.D.A.P.T.) using digital twin technology, reducing equipment downtime by 50%
- Built anomaly detection systems achieving 90%+ accuracy across 1M+ daily sensor events

Tata Consultancy Services (TCS) **India**
Software Engineer *Nov 2017 – Sep 2019*

- Automated QA testing with Selenium and Cucumber, improving testing efficiency by 70% across 10+ projects
- Developed AR application concepts shortlisted by TCS Innovation Labs for patent consideration

Research & Projects

Trust Bench | *LLM Trustworthiness Profiling Tool* **2026**

- Building tool for profiling LLM trust: extract internal signals, evaluate across safety dimensions, diagnose failures, and apply targeted improvements
- Studying trust signal behavior in Qwen3.5's hybrid attention architecture (linear vs full attention) using from-scratch implementation for full architectural access
- Integrating TrustLLM evaluation dimensions (truthfulness, safety, fairness, robustness) with mechanistic signal extraction

Predicting Risk of Lung Cancer From Medical History | *Quality Management in Health Care* **Apr 2025**

- Published peer-reviewed research achieving AUC 0.82 in lung cancer risk prediction using Electronic Health Records
- DOI: 10.1097/QMH.0000000000000525 | PubMed ID: 40197410

Production LLM Fine-tuning System | *Qwen2.5-3B with GRPO + LoRA* **2024**

- Fine-tuned 3B parameter model with 4-bit quantization and LoRA adapters, deployed via vLLM with auto-scaling, reducing inference costs by 25%

Education

George Mason University **Fairfax, VA**

Master of Science in Data Analytics *Aug 2022 – May 2024*

- Coursework: Machine Learning, Deep Learning, Statistical Analysis, Data Mining, Big Data Analytics
- Thesis: Advanced RAG Systems for Healthcare Applications

GITAM University **Visakhapatnam, India**

Bachelor of Technology in Electrical & Electronics Engineering *2013 – 2017*

Certifications

Google Cloud Professional Machine Learning Engineer | *Google Cloud* **2024**