Lucky Verma

667-910-4149 — luckie.verma30@gmail.com — linkedin.com/in/lucky-verma — github.com/lucky-verma — Nashville, TN

Professional Summary — Senior AI/ML Engineer with 5+ years specializing in production-scale agentic AI systems, multi-hop reasoning, and automated security analysis. Expert in end-to-end ML system design from research to deployment, with proven track record building enterprise solutions using LLMs, RAG architectures, and knowledge graphs. Passionate about solving complex problems through innovative AI applications.

Professional Experience

Eccalon LLC **June 2022 – Present**
Senior Machine Learning Engineer *Nashville, TN*

Advanced AI Systems & Intelligence Platforms

- Built AEGIS, a production-scale agentic AI platform combining LangGraph orchestration with DSPy optimization for complex multi-hop reasoning and foreign entity analysis
- Implemented sophisticated knowledge graph system using PostgreSQL + Apache AGE, enabling dynamic relationship discovery across 10M+ entities with 95% accuracy
- Developed professional Streamlit UI with FastAPI backend, serving 1000+ concurrent users with sub-10-second response times for complex analytical queries
- Architected enterprise-grade security including API authentication, rate limiting, and circuit breaker patterns for high-availability production deployment

Automated Security Analysis & Code Intelligence

- Designed CodeCompliant, an end-to-end automated security analysis platform integrating CodeQL static analysis with LLM-powered vulnerability assessment
- Built two-phase architecture supporting 8+ programming languages with automated remediation generation, reducing security review time by 85%
- Implemented Solr-based indexing system for efficient security finding search and retrieval across enterprise codebases
- Created autonomous code correction pipeline with automated PR generation using Git integration and LLM-powered fix suggestions

Computer Vision & Production ML Systems

- Developed Lirisyst, a production scene-text detection system using PyTorch with 95%+ accuracy for foreign language text processing and OCR integration
- Built GPU-accelerated 3D model optimization pipeline using CUDA and custom neural architectures, reducing processing time by 60%
- Implemented real-time sports analytics system with computer vision and deep learning for live game analysis, processing 30+ FPS video streams
- Deployed production RAG systems on AWS infrastructure serving enterprise data contextualization with 75% improved retrieval efficiency

University of Maryland, Baltimore County **January 2022 – May 2023**
Graduate Research Assistant *Baltimore, MD*

- Developed ensemble 1D U-Net architecture for biomedical signal classification using PPG data, achieving 92% accuracy on sleep-stage detection
- Researched real-time time-series processing for continuous health monitoring applications using signal processing and deep learning

Vast Dream Group **January 2021 – August 2021**
AI Specialist *Sydney, Australia (Remote)*

- Built production recommendation engine using BART for zero-shot classification, fine-tuned on MNLI dataset with 96% accuracy
- Architected AI measurement systems for optics industry, contributing to $1M+ revenue pipeline through automated quality analysis
- Designed and implemented end-to-end ML pipelines from data ingestion to model deployment using modern MLOps practices

Education

University of Maryland, Baltimore County **2023**
Master of Science in Computer Science – GPA: 3.97
Specialization: Machine Learning, Deep Learning, Advanced Algorithms

SRM University, Chennai, India **2019**
Bachelor of Technology in Electrical & Electronics Engineering – GPA: 3.9

Key Technical Projects

InfiniteContext-1B: Million-Token LLM System

- Built production-grade LLM system implementing DeepSeek-V3 Multi-Head Latent Attention (MLA) architecture with 1M token context window
- Achieved 93.7% KV cache memory reduction (8MB vs 128MB per 1k tokens) enabling inference on consumer GPUs (RTX 2070 Super for 128k context)
- Implemented distributed FSDP training on SLURM clusters with 92% GPU utilization on 4x A100s, custom Triton kernels for accelerated decoding
- Designed end-to-end MLOps pipeline with W&B experiment tracking, Kubernetes deployment via vLLM, and Ansible infrastructure automation

Multi-Modal RAG System with Real-time Web Integration

- Developed advanced RAG architecture combining document processing, web crawling, and vector search using ChromaDB and embedding models
- Implemented real-time content ingestion with SearXNG integration and automated fact-checking pipeline
- Built interactive knowledge graph visualization using PyVis with confidence scoring and source attribution
- Achieved 40% improvement in answer relevance over traditional RAG approaches through multi-path reasoning

Production MLOps Pipeline with Model Monitoring

- Designed end-to-end MLOps infrastructure using Docker, CI/CD, and automated model versioning
- Implemented comprehensive model monitoring with drift detection and automated retraining triggers
- Built centralized experiment tracking and hyperparameter optimization using modern ML tools
- Reduced model deployment time from days to hours through automated testing and validation pipelines

Technical Skills

LLM/GenAI	Fine-tuning, RLHF/DPO, RAG, Agentic AI, Tool Use, vLLM
Frameworks	PyTorch, LangChain, LangGraph, DSPy, HuggingFace
Distributed	FSDP, DeepSpeed, SLURM, Multi-GPU, Triton Kernels
Infrastructure	Kubernetes, Docker, FastAPI, PostgreSQL, Redis
Cloud/MLOps	AWS Bedrock, W&B, MLflow, CI/CD, Model Monitoring
Languages	Python, SQL, CUDA, C++, JavaScript, Bash
Vector DBs	ChromaDB, Pinecone, pgvector, FAISS, Weaviate
Specialized	Knowledge Graphs, CodeQL, Computer Vision, OCR

Recognition & Achievements

- Winner, HackUMBC 2021 – Best Docker Application, University of Maryland Baltimore County
- Led 100+ student technology organization as Vice Chairperson, managing 15+ technical projects and inter-college collaborations
- Published technical documentation and system architectures for enterprise-scale AI platforms
- Consistently exceeded performance targets, receiving top ratings across all technical competency areas