Venkatesh Shanmugam

Washington, DC • 7032162540 • venkatesh.shanmugam@gwu.edu

Founding AI/ML Engineer specializing in **production-grade generative AI** and **distributed ML systems** with **4+ years** scaling ML from startup to enterprise. **Pioneered Agentic AI solutions** achieving **90% query accuracy** and **40% cost reduction** while building **Graph-RAG systems** for healthcare applications. **Deep expertise in LLMs, federated learning, and clinical ML** with proven ability to deliver **78% prediction accuracy** and **sub-200ms latency** at scale. **Innovation leader** mentoring **100+ students** across **PyTorch, TensorFlow, Vertex AI, and Kubernetes** ecosystems.

WORK EXPERIENCE

ScriptChain Health — 05/2024 - Present

Founding Machine Learning Engineer — **Washington DC**

- Spearheaded end-to-end development and deployment of production-scale deep learning models for 30-day hospital readmission prediction, achieving 78% accuracy and enabling early clinical intervention at scale.
- Architected and implemented a distributed training pipeline with DeepSpeed and multi-GPU clusters, reducing epoch time by 67% (72 to 24 hours) and enabling scalable training for large models in healthcare workloads.
- Designed and streamlined CI/CD processes using Cloud Build, cutting model deployment timelines by 94% (from 4 hours to under 15 minutes) and delivering sub-200ms inference latency at 1,000+ RPS.
- Automated model promotion in Vertex AI with MLflow A/B testing, increasing PR AUC by 6 percentage points and reducing false positives by 8%, ensuring continuous production model improvement.
- Re-engineered Spark ETL pipelines on 500M+ records, reducing data preprocessing time by 99%, halving GPU usage, and saving $6,000 annually in infrastructure costs.
- Developed and launched a Graph-RAG recommendation system ("Food and Exercise as a Medicine"), increasing user engagement by 25% and recommendation relevance by 18% through advanced retrieval-augmented generation techniques.
- Demonstrated system design expertise by integrating ML lifecycle components, advanced MLOps practices, and scalable cloud-native architectures in fast-paced health tech environment.

The George Washington University — 08/2024 - 05/2025

Data Consultant • Part-time — **Washington DC**

- **Led technical mentorship and training initiatives**, designing and delivering 5 comprehensive workshops on machine learning, data visualization, and statistical methods to 100+ participants from diverse academic departments.
- **Provided advanced statistical consulting** to GWU academic community, supporting research projects with Python, R, SPSS, and STATA, ensuring methodological rigor and reproducible analysis workflows.
- **Expanded organizational capabilities** by introducing new analytical tools including GIS mapping and MATLAB, increasing team's technical versatility and research impact across multiple disciplines.
- **Architected knowledge transfer systems** by creating comprehensive tutorial materials and technical documentation, enabling self-service analytics adoption and reducing consultation dependency by 30%.
- **Facilitated cross-functional collaboration** between faculty, researchers, and graduate students, translating complex statistical concepts into actionable insights for non-technical stakeholders.
- **Demonstrated thought leadership** in applied ML and data science education, bridging academic research with industry-standard practices and modern data engineering workflows.

Tata Consultancy Services — 04/2021 - 08/2023

Software Engineer — **Chennai, India**

- **Architected and deployed enterprise-scale automation solutions** across 14+ client teams, reducing manual operational effort and cutting infrastructure costs by 40% through intelligent workflow optimization.
- **Designed and implemented data-driven process optimization systems** achieving 100% client adoption rate, directly aligning automated delivery pipelines with business KPIs and operational targets.
- **Engineered 12+ software automation bots** with seamless system integrations, improving process turnaround times by 30% and enabling scalable operations across multiple client environments.
- **Led cross-functional stakeholder engagement sessions** with technical and business teams, driving solution adoption strategies and measurably improving operational efficiency through collaborative problem-solving.

- **Demonstrated technical leadership** in enterprise transformation initiatives, managing complex client requirements while delivering robust automation frameworks that scaled across diverse industry verticals.
- **Established monitoring and analytics systems** for deployed automation solutions, enabling continuous optimization and identifying additional efficiency opportunities for sustained business impact.

EDUCATION

Master of Science in Computer Science

George Washington University • GPA: 3.78 Washington, DC • 09/2023 - 05/2025

Bachelor of Technology in Computer Science

SRM University • GPA: 3.5/4.0 Chennai, India • 08/2016 - 09/2020

PROJECTS

Agentic Graph RAG for Building codes

Architected production-scale multi-agent AI system with 5 specialized agents using LangGraph state machines and conditional routing, achieving 90% query accuracy improvement and 40% cost reduction.

Engineered high-performance dual-storage architecture (Neo4j + Redis) with asynchronous I/O operations, reducing query response time by 75% while supporting 1K+ RPS.

Implemented parallel processing pipelines with asyncio and concurrent execution, enabling 3x faster multi-modal document analysis and retrieval operations.

Pioneered transparent reasoning pipeline with multi-modal analysis and quality assurance scoring, reducing erroneous responses by 85% through intelligent decision-making.

Developed advanced retrieval-augmented generation with graph traversal algorithms and semantic similarity matching, improving information precision by 90%.

Built sophisticated agent communication protocols with tiered model orchestration and context-aware routing, enabling complex multi-step reasoning workflows.

Implemented comprehensive observability with LangSmith tracing, Prometheus metrics, and real-time monitoring dashboards, ensuring 99.9% system uptime.

AI-Text Discriminator

Pioneered advanced parameter-efficient fine-tuning techniques [PEFT] with Low-Rank Adaptation, reducing trainable parameters by **90%** while maintaining state-of-the-art classification performance.

Built a scalable NLP pipeline using LoRA-based fine-tuning of transformer models on a 1.2M+ text dataset, achieving **97%** accuracy with **3x** faster training and **50%** reduced memory usage.

Engineered hybrid feature extraction combining **semantic embeddings** and **syntactic patterns**, improving classification robustness and boosting real-world F1 score by **14%**.

Designed advanced feature engineering pipeline with **n-gram analysis**, **perplexity scoring**, and **attention pattern extraction** to capture AI-generated text signatures.

PUBLICATIONS

An End-to-End Unified Framework for Assessing Turning Movements Based Deep Neural Network

International Journal of Advanced Science and Technology

- Proposed a deep learning framework leveraging ANN, Kalman filter, and image processing to optimize urban traffic flow by accurately counting and forecasting turning movements under real-world constraints (occlusion, climate, dense traffic).
- Engineered a dynamic appearance model for robust vehicle tracking and introduced a path flow estimator for accurate hyperlink flow estimation, enabling improved city intersection design and traffic forecasting.

Enhanced Learning with Augmented Reality and Virtual Reality

International Journal of Engineering and Advanced Technology (IJEAT)

- Developed an AR-based platform for interactive visualization of biological diagrams, enabling students to scan organ images and access working AR models plus rich contextual information—improving engagement and comprehension in STEM learning.

Heart Beat Monitoring System and Alerting Indicator for Driver's Safety

International Journal for Science and Advance Research in Technology (IJSART)

- Designed and built an Arduino/IR-pulse-sensor system for real-time heartbeat monitoring in drivers, with embedded safety alerts to prevent accidents due to abnormal cardiac events—exploring IoT and embedded systems in health-tech applications.

Route Planning for Ships During Emergency Using Genetic Algorithm

International Journal of Emerging Technologies in Engineering Research (IJETER)

- Presented a genetic algorithm-based solution for optimal maritime route planning in emergencies (e.g., low fuel, ship damage), integrating ARPA system obstacle data and dynamic fitness functions to outperform classical pathfinding under constraint scenarios.

SKILLS

- **Programming Languages:** Python, SQL, C++, Java, R, Bash
- **ML/AI Frameworks:** TensorFlow, PyTorch, scikit-learn, Keras, LangChain, Transformers
- **Cloud & MLOps:** AWS (S3, Glue, Athena, EC2, SageMaker), GCP (Vertex AI, Cloud Build), Docker, Kubernetes (GKE), MLflow, CI/CD, Git, Jenkins
- **Data Engineering:** Apache Spark, Kafka, Airflow, Data Preprocessing, ETL Pipelines, Distributed Training, Model Monitoring, A/B Testing
- **Databases & Storage:** PostgreSQL, MongoDB, Redis, Pinecone (Vector DB), Neo4j (Graph DB), Elasticsearch
- **AI/ML Specializations:** Deep Learning, NLP, Computer Vision, Federated Learning, Agentic AI, Graph-RAG, Model Optimization (LoRA, PEFT)
- **Development Tools:** Linux, REST APIs, FastAPI, Streamlit, Jupyter Notebooks, VS Code, Weights & Biases, Prometheus, Grafana
- **Libraries & Utilities:** pandas, NumPy, Matplotlib, Seaborn, OpenCV, NLTK, spaCy, asyncio, Crew AI, Autogen