

Contact

vishnu71@colostate.edu

www.linkedin.com/in/vishnu71
(LinkedIn)

Top Skills

Amazon EKS

Unit Testing

ArgoCD

Languages

English (Professional Working)

Tamil (Native or Bilingual)

Telugu (Native or Bilingual)

Certifications

LLM Engineering: Master AI, Large Language Models & Agents

Problem Solving Basic

Business English Certificate Vantage

AWS Certified Cloud Practitioner (CLF-C02) Cert Prep: 2 Security and Compliance

Let's Talk

Honors-Awards

Master Of Ceremony

Student Achiever's Award

Best Paper Award - ICICICT International Conference

Certificate of Academic Excellence (for 4.0 GPA)

Publications

Deep Learning Techniques to Detect Learning Disabilities Among children using Handwriting (Scopus Indexed)

Vishnu Charan Venkatesh

AI/ML Engineer (SRE/DevOps) @ CCC | Ex-JP Morgan Chase |
Computer Grad @ Colorado State | Full Stack Software Engineer |
Agentic AI Automation Engineer
Seattle, Washington, United States

Summary

I'm an AI Solutions Architect, focused on building production-grade, agentic AI systems that operate reliably at scale. My work centers on designing end-to-end AI platforms using AWS, Kubernetes, and both proprietary and open-source LLMs, with an emphasis on low-latency inference, observability, and cost efficiency. I've led the development of real-time alert analysis and incident inference systems by integrating LLMs with cloud monitoring and on-call platforms, delivering significant improvements in response time, accuracy, and operational stability in mission-critical environments.

In parallel, I have deep hands-on experience fine-tuning and optimizing large language models, building RAG and graph-based knowledge systems, and implementing continuous evaluation and retraining pipelines to keep models accurate in production. Earlier in my career, I modernized large enterprise platforms by migrating monoliths to microservices and deploying cloud-native systems with CI/CD automation. I enjoy working at the intersection of AI, infrastructure, and reliability, turning advanced models into systems that deliver real, measurable business impact.

Experience

CCC Intelligent Solutions

AI Automation Engineer Intern

June 2025 - Present (8 months)

Chicago, Illinois, United States

- Built Agentic Alert Analysis system using AWS Bedrock LLMs, integrating PagerDuty, Prometheus, and AWS CloudWatch metrics, significantly reducing AI pipeline latency from 3 minutes to 50 seconds through async execution and LLM warm caching.
- Designed and integrated PostgreSQL (Amazon RDS) and Neo4J GraphDB as knowledge bases, optimizing query pipelines and graph traversal to enable

real-time alerting, achieving incident inference in under 200ms and 40% faster data retrieval.

- Implemented dynamic AI model switching, reducing cloud costs by 34.6% per quarter while maintaining performance SLAs.
- Built scalable evaluation pipelines using RAGAS and G-Eval to benchmark AI model accuracy, response consistency and reliability.
- Fine-tuned 13 open-source LLMs using PyTorch and HuggingFace, optimizing inference latency with vLLM and TensorRT-LLM to achieve 60% faster processing, enabling real-time analysis of 10,000+ incidents - improving overall system throughput in PROD.
- Designed retraining pipelines for LLMs with incremental learning on fresh alerts, improving model accuracy - 15% per quarter.
- Built RAG pipelines with LlamaIndex and Pinecone, deployed the Agentic system as Flask microservice on Amazon EKS.
- Earned recognition from the CTO and Global VP of P&T for demonstrating high-impact technical leadership in AI automation.

Colorado State University

AI Research Scientist

August 2023 - June 2025 (1 year 11 months)

Fort Collins, Colorado, United States

- Fine-tuned Llama 3.1-8B model using RAG with LangChain and Multi-shot prompting on National Weather data.
- Analysed model performance across Hugging Face LLM leaderboards, performed quantisation using Weights and Bias, generated vector embeddings using OpenAIEmbeddings, used Chroma and Pinecone for efficient vector storage.
- Fine tuned GPT 4o-mini using PEFT techniques, specifically QLoRA, achieving a 34% improvement in query handling efficiency compared to open source models.
- Developed a Weather Nowcasting application - radarca.engr.colostate.edu/public.
- Architected backend systems with Flask, created modular REST APIs for cross-platform integration and enhanced Database efficiency through schema optimisation and query tuning in PostgreSQL.
- Designed and deployed JWT token based auth with a session management system. Utilised docker to containerise application components and programmed Nginx for server health monitoring.

JPMorgan Chase & Co.

Software Engineer

September 2021 - August 2023 (2 years)

- Modernized 26+ enterprise applications from monolith to microservices, enabling modular integration of AI/ML components and improving inter-service latency by 15%. Tools - Spring Boot, REST APIs, Docker, Jenkins, PostgreSQL, Python, JUnit, Postman
- Built an auditing and analytics platform with PostgreSQL backend, automating data validation for 1K+ daily transactions and improving anomaly/error traceability. Tools - PostgreSQL, Splunk
- Developed technical documentation and deployment playbooks for AI microservices, enabling smoother onboarding and reducing deployment errors across teams. Tools - Confluence, Git

Education

Colorado State University

Master of Science - MS, Computer Engineering · (August 2023 - January 2025)

PSG INSTITUTE OF TECHNOLOGY AND APPLIED RESEARCH

Bachelor of Engineering - Computer Science and Engineering, Computer Programming/Programmer, General · (January 2018 - May 2022)