VISHNU MANOJ

Durham, NC | (919) 638 9588 | tm.vishnu.ms@gmail.com | linkedin.com/in/vishnu-mukundan-tm | calicartels.github.io

AIML engineer with experience building and deploying applied ML, computer vision, and LLM-based systems, including agentic workflows, with a focus on inference optimizations and production-ready pipelines. Proficient in delivering scalable, cloud-native AI solutions through academic projects, internships, and freelance work, using Python, PyTorch, FastAPI, and cloud platforms.

EDUCATION

Duke University — Durham, NC
Masters in Engineering, Artificial Intelligence; GPA: 3.72/4.0 — *Aug. 2024 – Dec. 2025*

- Relevant coursework: Sourcing for Data Analytics, Modeling Process and Algorithms, Business Fundamentals, Explainable AI, LLMs, Deep Learning, Optimization, Adversarial Attacks and Defenses

Vellore Institute of Technology — Vellore, India
Integrated Masters in Technology, CSE with Data Science; GPA: 8.32/10.0 — *Jul. 2019 – Jul. 2024*

- Selected coursework: Operating Systems, Data Structures, Analysis of Algorithms, Machine Learning, Data Science, Databases, Statistics, Mathematics, Linear Algebra, Business Intelligence, Predictive Modeling

TECHNICAL SKILLS

Languages: Python, Excel, JavaScript, HTML, SQL, Statistics
Frameworks: PyTorch, TensorFlow, Keras, NumPy, Pandas, PySpark, Scikitlearn, OpenCV, LangChain, LangGraph, HuggingFace
Platforms: Kali Linux, Windows, AWS, GCP, Microsoft Azure, Git, Docker, Kubernetes

EXPERIENCE

TRUST Lab Duke DeepTech — Aug 2025 – Present
Lead AI Developer — *Durham, North Carolina (On-Site)*

- Developed voice-based liaison agents as part of the research project in collaboration with OpenAI to explore how voice-based, conversational LLM agents can function as "research translators" in interdisciplinary collaborations.
- Orchestrated low-latency (800ms) text-to-speech agents using OpenAI Whisper, leveraging an orchestrator–worker workflow and agent memory for real-time information retrieval.

JPMorgan Chase & Co. — Jun. 2025 – Aug. 2025
AIML Associate Intern — *Plano, TX (On-Site)*

- Headed the automation of SAR Narrative by leveraging AWS bedrock and building a chain of thought workflow.
- Led automation of SAR Narrative generation using AWS Bedrock, cutting costs by $50,000 and reducing production time by 90% through prompt engineering.
- Also built a SHAP Explainer Model to provide explanations to outputs for combating anti-money laundering.

Sentics GmbH — Aug. 2022 – Nov. 2022
Computer Vision Intern — *Wolfsburg, Germany (Remote)*

- Engineered an algorithm that accurately estimated the base point of an object using pose keypoint data from TRTPose and 2D–3D correspondence, resulting in a 100% improvement in object location estimation accuracy.
- Conducted extensive research and experimentation with various object and keypoint tracking methods to evaluate performance trade-offs.

Miniscule Technologies Pvt. Ltd. — May 2022 – Jul. 2022
ML Cloud Deployment Engineer — *Chennai, India (On-Site)*

- Performed extensive research on evaluating major cloud service providers and their readiness for industrial 5G use cases as a Cloud AIOps Engineer.
- Deployed an on-edge custom face detection model through Amazon Rekognition trained on employee data stored on Amazon S3, achieving an accuracy of 88% on the Hikvision AcuSense camera module.

PROJECTS

PicoChat | *PyTorch, Gradio, HuggingFace, WandB, Tiktoken, NumPy* — 2026

- Used Nanochat by Karpathy as the teacher and distilled a 2B parameter language model to 375M parameters implementing Multi-Query Attention (12x memory reduction), multi-token prediction heads, and INT8 quantization (70.7% compression to 363MB); achieved 84% training loss improvement over 6000 steps on a single A100 GPU $9 total cost)
- Open-sourced complete training pipeline with reproducible commands, bug fixes, and documentation detailing distillation with MQA, multi-token prediction heads, and quantization techniques.

Duke Agentic Chatbot | *LangGraph, GCP, Pinecone, Claude API, Programmable Search Engines* — 2025

- Configured a GCP-native Modular Agentic chatbot that leverageed the orchestrator-worker workflow with the use of custom tools and internet search, with a latency of less than 15 seconds.

Blind.AI | *Python, YOLOv5, OpenCV, Twilio API* — 2023

- Built a voice- and gesture-based mobile application to assist individuals with acquired blindness.
- Implemented object detection via YOLOv5, currency detection using OpenCV, OCR through Pytesseract and an SOS signaling feature using the Twilio API.