Abbaas Alif Mohamed Nishar

Applied Scientist | LLM Engineering | Agentic AI | Multimodal AI | Computer Vision
470-601-4788 | abbaasalif6998@gmail.com | LinkedIn | GitHub | Google Scholar | Portfolio

Summary

Applied Scientist and Computer Science Ph.D. specializing in LLM fine-tuning, RAG, agentic systems, model evaluation, GPU serving, and multimodal computer vision. Built production-oriented conversational-AI systems, enterprise model-deployment platforms, and research prototypes spanning vision-language models, event cameras, optical communication, AI safety, and mechanistic analysis.

Core Skills

LLMs and Generative AI: Transformers, Hugging Face, PyTorch, PEFT, LoRA/QLoRA, TRL, RAG, constrained decoding, synthetic data, LLM-as-a-judge, test-time training, continual learning, ReAct agents, tool calling, vLLM, prompt engineering
Computer Vision and Multimodal AI: OpenCV, image segmentation, object detection, YOLO, GANs, vision-language evaluation, event cameras, optical camera communication, multimodal sensing, DALL-E inpainting
MLOps and Systems: Kubernetes, Helm, Amazon EKS, Kubeflow PyTorchJob, Docker, AWS S3, W&B, multi-GPU training/inference, tensor parallelism, FastAPI, React, CI/CD, model monitoring
Data and Research: Python, SQL, Pandas, NumPy, Scikit-learn, Snowflake, experiment design, ablations, bootstrap confidence intervals, human annotation, red-teaming, PyRIT, HarmBench, GFlowNets

Experience

Applied Scientist June 2025 – Present
Capital One – AI Foundations *San Jose, CA*
- Develop LLM-powered intent understanding, personalization, and agentic workflows for an enterprise conversational banking assistant.
- Fine-tuned Llama-based models with QLoRA across 300+ intents; combined retrieval, candidate-constrained decoding, and multi-turn context to improve long-tail accuracy by approximately 12 percentage points.
- Generated and self-verified 5,000+ synthetic examples across 200+ long-tail intents and evaluated retrieval, augmentation, reasoning, calibration, and confidence-gated inference strategies.
- Built evaluation pipelines for tens of thousands of conversational records and a React/FastAPI failure-analysis platform that exposed a systemic scoring mismatch and translated error patterns into model and product recommendations.
- Built a Kubernetes/Helm model platform for deploying Hugging Face models as GPU-backed vLLM services with S3 streaming, reusable recipes, validation, and an approval-gated operations copilot.

Founding Engineer July 2023 – June 2025
Revelio Communications Inc. *Atlanta, GA*
- Built a real-world screen-camera communication system for interactive television and mobile experiences; optimized GPU encoding by 100×.
- Developed a neural multi-stage mobile decoder that improved decodability by 60% while maintaining visually imperceptible video embedding.

Graduate Research Assistant Aug 2021 – June 2025
Georgia State University *Atlanta, GA*
- Conducted research in optical wireless communication, neuromorphic/event cameras, multimodal sensing, LLM-generated network simulations, traffic classification, and environmental time-series modeling.
- Built data-ingestion, visualization, and ML pipelines for asynchronous IoT, radon sensing, muon-based microclimate prediction, and data-agnostic image annotation.
- Published research across NLP, computer vision, optical wireless communication, and AI systems; received a Best Student Paper Award at IEEE SoutheastCon.

Data Science Intern May 2022 – Aug 2022
American Family Insurance *Remote*
- Developed image-segmentation models for home-inspection underwriting; reduced the training codebase by 80% and implemented automated multi-GPU training.
- Improved reproducibility and training throughput through modular experiment configuration and distributed execution.

DevOps Engineer Aug 2020 – Aug 2021
Tata Consultancy Services *Chennai, India*
- Improved retail forecasting accuracy by 25%, built operational dashboards, and developed Jenkins CI/CD pipelines that reduced deployment time by 50%.

Selected AI and LLM Projects

Enterprise LLM Intent Classification | *Llama, QLoRA, RAG, vLLM, PyTorch* 2026
- Built an 8B intent classifier across 300+ labels using multi-turn prompting, curriculum learning, vector-retrieved shortlists, and per-row constrained decoding; improved rare-intent accuracy by approximately 12 points.
- Created ambiguity audits, semantic evaluation, and synthetic-data pipelines to separate model errors from retrieval, taxonomy, and labeling limitations.

LLM Model Hub and Operations Copilot | *React, FastAPI, Kubernetes, Helm, vLLM* 2026
- Built a full-stack platform to download, configure, deploy, monitor, and troubleshoot GPU-backed LLM services across multiple GPU architectures and tensor-parallel configurations.
- Developed an approval-gated ReAct copilot with diagnostic tools, replay tests, deployment guardrails, conversation capture, and human-reviewed learning proposals.

Self-Adapting LMs and Continual Learning | *PyTorch, PEFT, LoRA, Multi-GPU* 2026
- Extended a self-adapting LLM framework with reinforced self-training, test-time adaptation, topic experts, multi-seed evaluations, and gradient-interference probes.

Diverse LLM Red-Teaming | *GFlowNets, PyRIT, LoRA, HarmBench* 2026
- Combined GFlowNet generation with a contextual bandit to produce diverse defensive red-team prompts; diagnosed reward misalignment and designed behavior-fidelity evaluation.

Mechanistic Analysis of LLM Refusal | *Transformers, Activation Probing, MoE Routing* 2026
- Used token-level activation trajectories, expert-router hooks, and causal ablations to identify generation-time refusal dynamics and distributed expert-routed behavior.

Education

Ph.D. in Computer Science May 2025
Georgia State University *Atlanta, GA*
M.S. in Computer Science May 2022
Georgia State University *Atlanta, GA*
B.Tech. in Electronics and Communication Engineering May 2020
Vellore Institute of Technology *Chennai, India*

Selected Publications and Patent

- **ART: Adaptive Reasoning Trees for Explainable Claim Verification** — Findings of EACL 2026.
- **Stress Tests REVEAL Fragile Temporal and Visual Grounding in Video-Language Models** — arXiv 2026; presented at two ICML 2026 workshops.
- **Revelio: A Real-World Screen-Camera Communication System with Visually Imperceptible Data Embedding** — IEEE ICASSP 2025.
- **Non-Line-of-Sight Optical Wireless Communication Using Neuromorphic Cameras** — EWSN 2025.
- **A Framework for Classifying Applications from Raw Network Traffic Traces** — IEEE SoutheastCon 2024; Best Student Paper Award.
- **Machine-Learning-Based System for Real-Time Screen-to-Camera Communication** — U.S. Patent Application 20250030811.

Leadership

President, IEEE at Georgia State University (2024–2025); organized an IEEE Atlanta ideathon, led a Raspberry Pi workshop with Girls Who Code, and delivered guest lectures in Robotics and Advanced Computer Vision.