Om Barde

New Delhi, India | ☎ (+91)8766721568 | ✉ bardeom6702@gmail.com
AI/ML | Generative AI | NLP | Computer Vision | PoC Development | Cloud & Big Data
https://ombarde.github.io/ | github.com/ombarde | https://www.linkedin.com/in/om-barde-622115171/

Profile Summary

- ➔ AI Developer with hands-on experience in **real-time video analytics and GPU-accelerated AI systems**.
- ➔ Strong expertise in designing and deploying **low-latency GStreamer-based video pipelines** for multi-camera RTSP/IP streams.
- ➔ Experienced in **training, fine-tuning, and deploying computer vision models** for vehicle detection, tracking, and event-based analytics.
- ➔ Proven ability to build **end-to-end AI solutions**, covering data preparation, model training, inference optimization, and system integration.
- ➔ Hands-on experience with **CUDA-enabled GPU acceleration** to optimize inference performance and system throughput.
- ➔ Practical exposure to **multi-GPU and multi-node AI systems**, including scalability testing and performance tuning.
- ➔ Direct experience executing **MLPerf Training and Inference benchmarks** on **NVIDIA H200 and L4 GPUs**.
- ➔ Strong understanding of **AI infrastructure**, including GPU compute, high-speed networking, storage, and system-level optimization.
- ➔ Skilled in analyzing **latency, throughput, GPU utilization, and memory performance** for AI workloads.
- ➔ Experienced in integrating AI services using **REST APIs** within Linux-based environments.
- ➔ Effective collaborator with cross-functional teams to align AI systems with **real-world operational and acceptance requirements**.

Professional Experience

AI Developer | CCS COMPUTERS PRIVATE LIMITED July 2025 - present

Project: GPU Benchmarking & Performance Evaluation

- ➔ Worked on **design, execution, and validation of large-scale GPU benchmarking** for AI compute and inference infrastructure as part of acceptance testing and deployment readiness.
- ➔ Performed benchmarking on **GPU-accelerated compute nodes (8× NVIDIA H200 per node)** and **inference nodes with NVIDIA L4 GPUs**, following **MLCommons MLPerf run rules**.
- ➔ Validated **training and inference performance** across multiple AI domains including **computer vision, NLP, speech recognition, recommendation systems, and LLM workloads**.
- ➔ Executed **single-node and multi-node MLPerf Training benchmarks**, including:

  3D-UNet (KiTS19) for medical image segmentation
  RetinaNet (OpenImages) and **Mask R-CNN (COCO)** for object detection
  ResNet (ImageNet) for image classification
  RNN-T (LibriSpeech) for speech recognition
  BERT (Wikipedia) and **DLRM-DCNv2 (Criteo 4TB)** for NLP and recommendation workloads
  GPT-3 (C4) and **Stable Diffusion v2 (LAION-400M)** for multi-node LLM and generative AI benchmarking

- ➔ Verified that **parallel benchmark instances** ran simultaneously on the complete system while **individually meeting MLPerf timing and throughput criteria**.
- ➔ Conducted **MLPerf Inference (99% Offline) benchmarking**, validating throughput targets for:
- ➔ **3D-UNet, RetinaNet, ResNet, BERT, RNN-T, GPT-J, and DLRM-v2**
- ➔ Ensured compliance for both **single-node inference** and **per-GPU (L4) inference targets**
- ➔ Analyzed performance across **precision modes (FP32, FP16, FP8, INT8)** to evaluate compute efficiency and model scaling behavior.
- ➔ Assessed **GPU utilization, memory bandwidth, interconnect performance, and latency** under high-concurrency workloads.

- ➔ Documented **benchmark methodology, results, and acceptance reports**, ensuring results were **well-documented and countersigned** as per acceptance testing requirements.

Project: Speed Detection & Monitoring System

- ➔ Implemented a **real-time, GPU-accelerated vehicle speed detection system** for traffic monitoring using **GStreamer-based video pipelines**.
- ➔ Built **end-to-end GStreamer pipelines** for ingesting live **RTSP/IP camera feeds**, handling decoding, color conversion, buffering, batching, and frame synchronization with minimal latency.
- ➔ Designed and **trained deep learning–based object detection models (YOLO variants)** on traffic surveillance datasets to accurately detect vehicles under varying camera angles, lighting, and traffic densities.
- ➔ Performed **dataset curation, annotation validation, and preprocessing**, including resizing, normalization, and data augmentation to improve model robustness.
- ➔ Fine-tuned trained models for **real-time inference**, balancing accuracy and latency for production deployment.
- ➔ Integrated trained models with **OpenCV-based post-processing** for bounding box filtering and object association.
- ➔ Developed **object tracking logic** across consecutive frames to maintain vehicle identity over time.
- ➔ Implemented **speed estimation algorithms** by computing frame-to-frame displacement, correlating timestamps, and applying camera calibration and perspective correction.
- ➔ Leveraged **CUDA-enabled GPU inference** to accelerate trained model execution and minimize CPU–GPU memory transfer overhead.
- ➔ Tuned **GStreamer elements (queue sizes, buffering strategy, batch size, frame skipping)** to balance detection accuracy and real-time throughput.
- ➔ Implemented **event detection logic** to identify over-speeding vehicles and trigger alerts, logs, and downstream reporting.
- ➔ Validated trained model performance and end-to-end system behavior across **different lighting conditions, camera placements, vehicle densities, and environmental scenarios**.
- ➔ Ensured long-running pipeline stability by handling **stream drops, reconnections, and runtime pipeline recovery**.

Project: Dock Management System (DMS)

- ➔ Developed an **AI-driven Dock Management System** to automate monitoring of vehicle movement, dock occupancy, and turnaround time in large-scale logistics facilities.
- ➔ Designed **multi-camera GStreamer pipelines** to ingest, synchronize, and process real-time video feeds from dock entry, exit, and loading/unloading zones.
- ➔ Designed and **trained computer vision models** for **vehicle detection and classification**, enabling differentiation between trucks, containers, and operational vehicles.
- ➔ Prepared and curated **domain-specific datasets**, including annotation verification and class balancing to improve model accuracy in crowded dock environments.
- ➔ Trained the **Tr-OCR** model on custom Datasets to fetch the INDIAN Truck Number plates.
- ➔ Fine-tuned trained models to handle **partial occlusions, overlapping vehicles, and varying operational layouts**.
- ➔ Implemented **ROI (Region of Interest) and zone-mapping logic** to associate detected vehicles with specific docks and operational areas.
- ➔ Developed **state-based vehicle tracking** to identify entry, docking, idle, loading/unloading, and exit events.
- ➔ Integrated trained AI inference outputs with backend services via **REST APIs** to update dock occupancy status, trigger workflows, and populate operational dashboards.
- ➔ Leveraged **GPU acceleration** to scale inference as the number of docks and concurrent video streams increased.
- ➔ Implemented **event-driven analytics**, including:

  - ◆ Dock utilization time
  - ◆ Vehicle wait and turnaround time
  - ◆ Congestion and bottleneck detection
  - ◆ Ensured system robustness by handling **camera failures, stream interruptions, and complex vehicle movement patterns**.

- ➔ Collaborated closely with operations and infrastructure teams to align trained AI models and system outputs with **real-world dock workflows and constraints**.

Data Analyst | UNIFINS IT HUB Jun 2024 – July 2025

Project: Multilingual Sentiment Classifier using BERT

- ➔ Designed and deployed a **multilingual sentiment analysis engine using BERT**, achieving over 90% accuracy across English, Hindi, and Marathi feedback.
- ➔ Integrated model into internal CRM and analytics platforms via **RESTful APIs** using **GCP Cloud Run**, enabling real-time sentiment classification at scale.
- ➔ Built modular **ETL pipelines** in **Python** (**Pandas**, **spaCy**) to process and clean structured/unstructured feedback data; stored results in **BigQuery and Google Sheets**.
- ➔ Used **RAKE** to extract dominant keywords and recurring complaint patterns from user feedback to support trend analysis.
- ➔ Delivered AI-powered features such as **feedback categorization, sentiment tagging, and priority routing** to enhance support response efficiency.
- ➔ Led a 3-member team and collaborated with frontend/backend developers to ensure smooth deployment and system integration.

Project: Unifins Wealth Management- Model Research

- ➔ Led a **comparative market study** of platforms like AssetPlus, Groww, and Zerodha Coin to design a scalable, SEBI-compliant mutual fund distribution model with insights on **commission structures**, **AMC partnerships**, and **client onboarding workflows**.
- ➔ Mapped onboarding, KYC/CKYC, and transaction workflows using **Lucidchart**, and documented standard operating procedures using **Notion**, ensuring regulatory alignment and **process clarity**.
- ➔ Designed internal operations by defining **role hierarchies**, team responsibilities (RM, Support, Compliance), and built a draft **SOP manual** to streamline execution.
- ➔ Created a lightweight CRM using **Google Sheets and Google Forms** to track client data and fund performance, supporting **data centralization** and informed decision-making.
- ➔ Developed a **go-to-market strategy** including investor education modules and onboarding kits for distributors, enhancing outreach and conversion efficiency.

Research & Development Intern | University of Texas at Austin | Remote Jun 2023 – Dec 2023

- ➔ Worked on the estimation of **object dynamics from video sequences**, applying **deep learning models** with PyTorch and OpenCV.
- ➔ Collaborated with the professor to address the problem of **estimating object dynamics and motion patterns from unstructured video data**, relevant to applications like surveillance, autonomous systems, and robotics. Gained exposure to distributed AI research practices and global collaboration.
- ➔ Developed and trained deep learning models using **PyTorch** and **OpenCV** to analyze video sequences and predict temporal object movement.
- ➔ Built preprocessing pipelines to extract and annotate frames, applied **CNNs for spatial features**, and integrated **RNNs (e.g., LSTM)** to capture sequential dependencies for motion prediction.
- ➔ Conducted **iterative optimization and performance tuning** to improve accuracy and temporal coherence of tracking.

AI/ML INTERN | The Sparks Foundation | Remote Feb 2023 – Mar 2023

- ➔ Built supervised learning models using Scikit-learn, including classification and regression with real-world datasets.
- ➔ Applied visualization and statistical analysis for feature exploration and model evaluation.

AI INTERN | TechKrow | Remote Nov 2021 – Dec 2021

- ➔ Assisted in the design and deployment of basic ML systems; implemented preprocessing pipelines and predictive models.

Skills

- ➔ **Programming Languages:** Python, C++, C, SQL, NoSQL, R
- ➔ **AI / Machine Learning:** Machine Learning, Deep Learning, Computer Vision (Detection, Tracking, Video Analytics), Transformers, LLMs (BERT – training & inference), Time-Series & Event-Based Analytics, Model Training, Fine-Tuning & Optimization, Data Preprocessing & Dataset Curation
- ➔ **Video Analytics & Streaming:** GStreamer (RTSP pipelines, multi-stream processing, buffering & batching), Real-time Video Processing Pipelines, OpenCV
- ➔ **GPU, Acceleration & Benchmarking:** CUDA Programming (Foundations), GPU-Accelerated Training & Inference, Multi-GPU & Multi-Node Scaling, Precision Optimization (FP32, FP16, FP8, INT8), GPU Performance Profiling & Optimization, MLPerf Training & Inference Benchmarking
- ➔ **Frameworks & Libraries:** PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, Pandas, NumPy, Matplotlib
- ➔ **Software Development:** Object-Oriented Programming (OOP), Design Patterns, SDLC, Data Structures & Algorithms, REST API Development & Integration, Debugging & Performance Analysis (GDB, Valgrind), Build Systems (CMake, Makefiles)
- ➔ **Backend, Cloud & DevOps:** Linux (Ubuntu, RHEL), Docker (Fundamentals), Google Cloud Platform (GCP), Cloud Deployment & API Integration, CI/CD (Basic), Version Control (Git)
- ➔ **AI Infrastructure & Systems:** NVIDIA GPUs (H200, L4, A5000), GPU Server & AI Infrastructure Optimization, High-Throughput AI Systems, Acceptance Testing & Performance Reporting
- ➔ **Professional & Research Skills:** Agile Development, System Design & Optimization, Performance Tuning, Technical Documentation & Reporting, Cross-Team Collaboration, Research Writing & Publication

Education

G.H. RAISONI INSTITUTE OF ENGINEERING AND TECHNOLOGY, NAGPUR
Bachelor in Technology| Artificial Intelligence 2020 – 2024

- CGPA: 8.25/10

Publications

"SatelliteChangeNet: Deep Learning approach for Detection & Prediction"
International Journal of Advanced Research in Science, Communication and Technology, May 2024

Certifications

- ➔ **Compute Technical Curriculum** – NVIDIA
- ➔ **NVIDIA AI Technical Curriculum** – NVIDIA
- ➔ **NVIDIA DGX Technical Curriculum** – NVIDIA
- ➔ **Networking Technical Curriculum** – NVIDIA
- ➔ **Advanced Prompt Engineering Techniques** – LinkedIn
- ➔ **Career Essentials in GitHub Copilot (Professional Certificate)** – GitHub
- ➔ **Docker Essentials: A Developer Introduction** – IBM
- ➔ **Fundamentals of Deep Learning** – NVIDIA
- ➔ **Accelerating End-to-End Data Science Workflows** – NVIDIA
- ➔ **Getting Started with AI on Jetson Nano** – NVIDIA

- ➔ **Disaster Risk Monitoring Using Satellite Imagery** – NVIDIA
- ➔ **Develop, Customize, and Publish in Omniverse with Extensions** – NVIDIA
- ➔ **Google Cloud Professional Cloud Architect** – Udemy
- ➔ **Python Data Science Toolbox** – DataCamp
- ➔ **Google Analytics Certification** – Skillshop
- ➔ **Entrepreneurship** – NPTEL
- ➔ **Graphic Design: Layout & Composition** – LinkedIn Learning
- ➔ **Certified Web Designer Associate (CWDSA)** – MKCL

Key Projects

OBJECT DYNAMICS ESTIMATION FROM VIDEO (UT AUSTIN PROJECT)

- Collaborated remotely on a research project under **Prof. Chandrajit Bajaj**.
- Developed models to estimate object motion and predict physical dynamics from real-world video footage.
- Focused on temporal pattern recognition and movement estimation using CNNs + RNNs.
  Impact: Aided ongoing research on motion understanding and physical scene interpretation.

SATELLITECHANGENET – DEEP LEARNING FOR SATELLITE IMAGERY

- Developed an AI system to detect changes in satellite imagery using deep learning architectures.
- Benchmarked **U-Net** (85% accuracy for fine-grained segmentation), **YOLO** (88% real-time object detection), and **R-CNN** (83% for localization).
- Designed modular architecture and followed SDLC phases including requirement analysis, design, and testing.
- Published results in **IJARSCT (2024)**.
  Impact: Enabled improved land-use monitoring and disaster analysis.

STOCK PRICE PREDICTION USING LSTM

- Designed and trained a deep learning model using **LSTM networks** to predict stock prices.
- Applied **backpropagation through time** and **stochastic gradient descent** to update weights. Conducted time-series analysis and evaluated MSE/RMSE for multiple time horizons. Also, designed reusable code structure for extensibility and future model integration.
  Impact: Demonstrated feasibility of AI for stock forecasting under noisy financial data.

ROBOTIC PATH PLANNING

- Implemented pathfinding algorithms using both **Holonomic and Non-Holonomic movement strategies**.
- Developed simulation models and decision systems for robot navigation in static environments.
- Visualized paths and optimized for shortest route using geometric heuristics.
  Impact: Contributed to foundational logic for autonomous navigation and motion planning.

Leadership & Volunteering

- **Vice Chair**, IEEE SB GHRIETN
- **President**, AIBOTRIX – AI Forum
- **Member**, IEEE India
- **Volunteer**, Feel Good Foundation & I-SMART India

Languages

- English – Professional Proficiency
- Hindi & Marathi – Native Proficiency