Ayush Dongre

(+91) 9623704912
dongreayush39@gmail.com

Machine Learning Engineer

github.com/k0y0min
linkedin.com/in/ayush-dongre

Education

Indian Institute of Technology, Kharagpur **2019 — 2024**
Integrated M.Sc. (5-Year) in Mathematics & Computing *CGPA: 7.07/10*

Experience

Independent ML Research & Developer **Jun'25 - Present**
GraphRAG | Hosted via Google Cloud Run | Live

- Architected a serverless RAG pipeline, minimizing idle infrastructure costs via on-demand scaling of **Dockerized** microservices.
- Reduced indexing costs by **90%** relocating segmentation to local SLMs (**Qwen 3 : 0.6B**) & using SOTA APIs for entity extraction.
- Implemented **"Ancestry Injection"**, embedding document hierarchies, solving context fragmentation & **lost in the middle** problem.
- Replaced imprecise chunking with **perplexity-based segmentation** using dynamic log-loss thresholds to split at narrative breaks.
- Integrated **KuzuDB** to enable **O(1) adjacency traversals** & complex **Cypher** queries, replacing unbundled Parquet-based storage.

Data Scientist, Turing **Jun'24 - May'25**
USA | Remote

- Evaluated models using **Python** & **C++** red-teaming, ensuring strict instruction-following and factual correctness confirmations.
- Built **SFT datasets** with GitHub repos to fine-tune **Gemini 1.5 Pro** for code, ensuring robustness via Docker-based unit testing.
- Developed Python pipelines to generate 200+ complex data visualisations that trained **Gemini's** multimodal capabilities efficiently.
- Audited models via RLHF on 2K+ math & 200+ coding problems, finding logical & stylistic flaws to improve response reliability.

Global Data Insights and Analytics Intern, Ford Motors **May'23 - Jul'23**
Chennai, India

- Engineered **NLP classification pipelines** to automate high-volume ticket routing, greatly reducing manual triage complexity.
- Fine-tuned **MiniLM** models via **UMAP** & **HDBSCAN** clustering to automate labeling, achieving sustained **85% balanced accuracy**.
- Migrated legacy ETL to production-grade **Python** pipeline, optimizing logic to reduce manual approval intervention by over **30%**.

Research Intern, IIM Mumbai **Nov'22 - Mar'23**
Mumbai, India

- **Published** an article on 'Knowledge Graph Driven **Credit Risk Assessment**' in the International Journal of Production Research.
- Curated & refined a dataset by extracting MSME data from the CMIE database for over **5000** MSMEs spanning a period of 6 years.
- Implemented **RGCN** & **TransE** to generate 50-dimensional robust vector embeddings from the synthesized knowledge graph.
- Achieved a balanced accuracy of **92%** on the selected model (RGCN-RF), significantly outperforming major benchmark models.

Projects

Seam Carving | Content-Aware Image Resizing

- Engineered high-performance **C++/CUDA** tool to achieve **7x speedup** vs. Python with parallel dual-gradient energy computation.
- Implemented custom **CUDA kernels** and **row-wise parallelized DP** to calculate cumulative energy maps across the GPU threads.
- Utilized OpenCV's `cv::cuda::GpuMat` for device memory, reducing host-device transfer overhead for batch input image load.

Reliability Analysis of Neuro Evolution using Conformal Prediction
Master's Thesis Project

- Delivered **risk-aware** neural networks that adapt confidence intervals to minimize uncertainty for safety-critical applications.
- Augmented **NEAT** fitness function with **Conformal Prediction**, evolving architectures via peak **average prediction set size** metric.
- Achieved **>96% accuracy** on Iris/XOR evolving minimal topologies using 3/4 inputs while ensuring marginal statistical coverage.
- Validated on complex **Letter-Recognition** (20k instances) achieving convergence in **460 generations** via custom mutation rates.

Technical Skills

Languages	C/C++, Python, Bash, Assembly, SQL
Libraries/Frameworks	Pytorch, Keras, Sklearn, OpenCV, PyKeen, TensorFlow
Developer Tools, Concepts	Docker, Nvim, Git, CUDA, CI/CD, Unit Testing
Relevant Coursework	Design and Analysis of Algorithms, AI: Foundations and Applications, Object Oriented Systems, Computer Networks, Switching and Finite Automata, Graph Theory and Algorithms, Non-Linear Programming

Certifications

- NVIDIA Deep Learning Institute Certifications:
  Accelerated Computing with CUDA C/C++, Transformer-based NLP Applications, and Fundamentals of Deep Learning.