Dhanesh Baalaji Srinivasan

ds7636@nyu.edu | +1 9292932179 | Linkedin | Github | Seattle, WA (Open to relocate)

EDUCATION

New York University, New York City, NY; Master of Science in Computer Science; **GPA**: 3.75/4; **Honors*:*** Merit-based scholarship

TECHNICAL SKILLS

Programming Languages: Python, C, C++, C#, JavaScript, SQL. **Frameworks:** Django, .NET Core, Angular, PyTorch, JAX, TensorFlow.
ML Systems: Retriever-Augmented Generation (RAG), LLM Fine-tuning (QLoRA, Unsloth), IBM AnalogNAS, IBM AIHWKit, CUDA, Triton.
Cloud & DevOps: AWS, Docker, Kubernetes, CloudFormation. **Distributed Systems & Big Data:** Spark, Hadoop, MapReduce, Cassandra.
Databases & Search: OpenSearch, Elasticsearch, PostgreSQL, MongoDB, SQL Server. **Tools:** Linux, Git, Postman, Visual Studio.

WORK EXPERIENCE

New York University, New York City, United States Jul 2025 - Present
Software Development and DevOps - Project Lind, NYU Secure Systems Lab | Python, Docker, C
- Refactored a Python test runner to prevent inconsistent execution states by eliminating unsafe system calls, improving reliability.
- Redesigned test setup with a temp directory and auto-cleanup; removed redundant file operations, reducing I/O overhead by 70%.

LOCOMeX, Inc., New York City, United States (Remote). Feb 2025 - May 2025
Software Engineer and MLOps Intern | Django, AWS Lambda, DynamoDB, RDS, Postgres, Python, Tailwind CSS, Bootstrap.
- Engineered a full-stack, low-latency autocomplete feature using AWS Lambda and RDS, improving search responsiveness by 70%.
- Built Django APIs and containerized ML models as serverless Lambda functions, enabling scalable, low-latency inference.
- Built a PDF reporting tool with visual analytics in Django, reducing manual report generation time from 1 hour to under 1 minute.

New York University, New York City, United States Jan 2024 - May 2025
Graduate Research Assistant - Brooklyn Application,Architecture,Hardware Lab| DARPA Project | C, Python, Assembly, ARM NEON
- Integrated a Last-level Cache into a Spectrum sensing Processor simulator and created sweeps to obtain the optimal cache size.
- Modeled and introduced variable Common Bus delays to assess signal detection throughput under various latency constraints.
- Developed Power Spectral Density and Match filter kernels using ARM v8.2 NEON for real-time spectrum sensing computations.

Graduate Course Assistant - High Performance Machine Learning and Big Data | Pytorch, CUDA, C, Pyspark, Dask, MapReduce
- Created and graded Homework Assignments for two graduate courses in Pytorch, CUDA, C, MapReduce, MongoDB and Spark.
- Assisted the Professors in developing Lecture materials on topics like CUDA, Distributed Training, and Apache Cassandra.

Psiog Digital Private Limited, Chennai, India Nov 2020 - May 2023
Software Engineer | Angular, ASP .NET Core, ASP .NET Framework, ASP .NET MVC, Javascript, C#, SQL.
- Devised a scalable Bidding system that incorporated an AI voice assistant which generated 4000+ user registrations within a month.
- Crafted RESTful APIs, designed SQL scripts, and responsive User Interfaces resulting in a 30% increase in user engagement.
- Fixed critical Extract, Transform, and Load (ETL) pipeline issues, saving $200k in potential losses from data downtime.
- Managed CI/CD pipelines across multiple products, reducing deployment times by 25% and hence improving release frequency.

PROJECTS

LlamaLearn - Retriever-Augmented Generation (RAG) flow in AWS for Large Language Models (LLMs) | Amazon Web Services, Python.
- Architected a scalable RAG system using DPR for dense retrieval and NeuralHermes-2.5 (Mistral-7B) for generation, deployed via AWS EKS and ECR with OpenSearch for vector search and DynamoDB for user-specific metadata to enable personalized answering.
- Engineered a modular information retrieval pipeline featuring document chunking, DPR-based vectorization, and OpenSearch k-NN search, enabling low-latency, semantically accurate real-time question answering.
- Improved answer quality and reduced hallucinations by injecting top-k retrieved chunks into the LLM for context-aware generation.

Fine-tuned Llama 3.1 8B for Math Question Answering | **Deep Learning** | Pytorch, Numpy, unsloth, huggingface
- Fine-tuned LLaMA 3.1 8B for math question answering using Rank-Stabilized LoRA and structured prompt engineering, achieving 82.04% test accuracy which is a 9.4% relative improvement over the 75% baseline.
- Leveraged 4-bit quantization and Unsloth's memory-optimized training stack to fine-tune LLaMA 3.1 8B on a single T4 GPU.

NAS-SegNet - A Novel efficient neural network for Medical Image Segmentation | NYU and IBM | PyTorch, IBM AnalogNAS, AIHWKIT
- Designed NAS-SegNet, an 800K-parameter model achieving 0.58 IoU digitally, matching U-Net performance with 90% fewer parameters, and 0.40 IoU under analog noise using AIHWKit, validating deployment on analog accelerators.
- Modified IBM's classification supernetwork for segmentation by adding transpose-convolution layers for pixel-wise prediction.

Subreddit Recommendations and Sentiment analysis on Reddit data | Pyspark, DistilBERT, TF-IDF
- Analyzed 3.8M Reddit posts using PySpark, TextBlob and DistilBERT to perform large-scale sentiment classification.
- Developed a content-based subreddit recommendation system using TF-IDF vectorization and cosine similarity to rank subreddits by the volume of posts exceeding a semantic similarity threshold with the user query.