

# BADRI NARAYANAN MURALI KRISHNAN

✉ mbadriwork@gmail.com

☎ +1 848 - 363 - 8431

🌐 linkedin.com/in/mbadrinarayanan

🐙 github.com/MBadriNarayanan

🌐 mbadrinarayanan.com

📍 San Diego, California

## SUMMARY

**AI/ML Engineer** with an **MS in Computer Science** delivering production-grade **GenAI / ML solutions** from **0 to 1**. Experienced in **ML model optimization**, **Multi-Agent orchestration**, **RAG Pipelines**, and **LLM fine-tuning**, with a proven track record of translating complex business requirements into robust **AI systems**. A published researcher (**ICLR 2025**) with deep technical expertise in fine-tuning, quantization, and safety-critical AI design across major cloud platforms and edge devices.

## TECHNICAL SKILLS

**GenAI Competencies:** Retrieval-Augmented Generation (RAG), Agentic AI Systems, Multi-Agent Orchestration, Prompt Engineering, Structured Outputs, Tool Calling, NLP

**LLM Expertise:** Large Language Models (LLMs), LLM Fine-tuning (QLoRA), Instruction Tuning, Model Evaluation, Machine Learning

**LLMs & Model Ecosystem:** GPT-4, Gemini Pro/Flash, Llama, Mistral, T5, BERT, Local LLMs (Ollama)

**Agentic Frameworks & GenAI Tools:** LangGraph, LangChain, CrewAI, LlamaIndex, Hugging Face Transformers, Chain-of-Thought Reasoning, Vector Databases (Chroma, FAISS, Pinecone), Weights & Biases

**Cloud & AI Platforms:** Google Vertex AI, AWS SageMaker, Azure AI Foundry, Azure AI Studio, GCP, AWS, Azure

**Backend & APIs:** Python, C, C++, FastAPI, RESTful APIs, Postman, Microservices Architecture, Async Processing

**ML Frameworks & Data Systems:** PyTorch, TensorFlow, Spark, SQL, BigQuery, ETL Pipelines

**Observability & MLOps:** Prometheus, Grafana, Model Evaluation & Benchmarking, Experiment Tracking, CI/CD

**DevOps & Deployment:** Docker, Git, Cloud-native Deployment, Scalable Inference, Model Serving

**Libraries & Visualization:** NumPy, Pandas, Scikit-learn, Matplotlib, Streamlit

## WORK EXPERIENCE

### Machine Learning Engineer

March 2026 – Present

Qualcomm

San Diego, CA

- Working to design and **optimize** high-performance **ML and GenAI models** on **ONNX Runtime** for optimal performance on **Snapdragon SoCs**.
- Applying **advanced optimization techniques**, including **model quantization** and **graph-level transformations**, to ensure efficient execution and high system reliability on power-constrained hardware.

### AI / ML Engineer

September 2025 - February 2026

Ergo, USA, Inc

Madison, WI

- Founding AI Engineer** working to build intelligent agentic systems that enhance the platform's **Agentic AI capabilities**.
- Architected a **Medication Agent** combining **LLM prompts** with **structured data parsing** and **fallback handling** to ensure reliable medication safety insights.
- Built a **centralized AI caching microservice** using **FastAPI** that reduced duplicate LLM API costs by **40%** and latency by **25%**.
- Powered production LLM workflows using **Google Vertex AI** and implemented end-to-end **observability** with **Prometheus and Grafana** to monitor latency, cost, and system health.

### AI / ML Engineer

June 2025 - September 2025

BrainWaves Digital LLC

Remote, USA

- Lead engineer in creating a **Bank Statement Processing Engine** from **scratch** to extract financial data & insights.
- Outlined **schemas** detailing extraction features, used **few-shot prompting** to categorize transactions into buckets.
- Engineered a **multi-model AI stack** comprising **Gemini 2.5 Pro/Flash, GPT-4** with intelligent fallback mechanisms, improving extraction accuracy by **35%** and reducing manual review time by **44%**.
- Implemented a frontend analytics dashboard with **Supabase SSO Login** and **Firestore integration**, visualizing cash flow trends, balances, and categorized income/expenses.
- Created a dashboard with **ReactJS, Supabase SSO & Firestore** to visualize the cash flow, balances, & transactions.

## Graduate Research Assistant

September 2024 - May 2025

University of Wisconsin – Madison

Madison, WI

- Designed a scalable ingestion and processing system for **80k+** medical documents, optimizing chunking, embedding, and retrieval efficiency.
- Developed an **Agentic RAG** integrating retrieval, reasoning, and multi-agent orchestration to provide contextualized diagnostic insights for radiology.
- Generated document embeddings and stored them in **Chroma DB**, serving locally hosted LLMs via **Ollama** for low-latency experimentation.
- Established an evaluation framework using **NDCG@5**, achieving a score of **0.857** on 1K+ radiology reports.
- Benchmarked and profiled multiple **Large Language Models (LLMs)** under diverse configurations and hardware settings to analyze performance, latency, and efficiency trade-offs.
- Investigated LLM **compression** and **quantization techniques** to optimize deployment on edge and mobile platforms without compromising accuracy.
- Assessed model safety and quality trade-offs**, measuring accuracy, hallucination frequency, and toxic output in compressed LLMs.
- Co-authored PalmBench, accepted at **ICLR 2025**, on compressed LLM evaluation for mobile use cases.

## Autonomous System Research Intern

June 2024 - August 2024

Nokia Bell Labs

Murray Hill, NJ

- Conceptualized a system to explore the use of **autonomous agents** for problem-solving, applying them to a controlled environment to evaluate adaptive learning and collaborative reasoning.
- Created a multi-agent framework where each **agent** performed specialized actions through **tool-calling interfaces**.
- The agents were coordinated by a central **Orchestrator LLM**, which planned & optimized multi-step task execution.
- Implemented the **multi-agent system** using **LangGraph** & **Chain-of-Thought** prompting to optimize reasoning and decision quality across agents.
- Deployed the system via **Streamlit** for interactive visualization and analysis of multi-agent behaviors, enabling real-time testing and iterative refinement.

## Associate Engineer

July 2022 - July 2023

Qualcomm

Hyderabad, India

- Introduced evaluation frameworks and metrics for ML models to enhance efficiency and accuracy by **10.26%**.
- Benchmarked quantized and pruned models on Snapdragon SoCs to assess latency, throughput, and power consumption under production workloads on SNPE (SnapDragon Neural Processing Engine).
- Collaborated with hardware and firmware teams to fine-tune AI workloads for mobile deployments, ensuring consistency across diverse device configurations.

## SELECTED PROJECTS

---

### Relationship Extraction using Large Language Models [Repo](#)

- Fine-tuned **Llama3.2 (8B) LLML** for end-to-end relation extraction using the REBEL framework, adapting the sequence-to-sequence approach to generate structured triples (Entity 1, Relationship, Entity 2).
- Engineered a full data pipeline to preprocess REBEL datasets for LLAMA input and executed multi-epoch fine-tuning to capture all possible relational patterns.
- Evaluated and benchmarked performance against the REBEL baseline, analyzing Precision, Recall, and F1 scores to assess generative LLM strengths and limitations for structured information extraction.

### Optimizing Natural Language Understanding: Fine-tuning Mistral 7B [Repo](#)

- Fine-tuning **Mistral 7B LLM** on the Samantha dataset for conversational dataset for instruction-following and contextual Q&A.
- Implemented a **QLoRA** fine-tuning pipeline with **4-bit quantization** using Hugging Face TRL, PEFT, and bitsandbytes.
- Integrated **W&B** logging for training metrics and checkpoints, and developed inference scripts for adapter-based response generation.

### Multi-GPU Distributed Training of Large Language Models using Megatron-LM [Repo](#)

- Designed and implemented a **multi-GPU training pipeline** for large language models using **Megatron-LM** integrated with **Hugging Face Accelerate** to enable efficient distributed fine-tuning.
- Enabled scalable **data, tensor, and pipeline parallelism** for transformer models, improving training throughput and GPU utilization across multi-node environments.
- Built end-to-end workflows for **LLM fine-tuning, checkpointing, and evaluation** on distributed hardware, ensuring reproducibility and efficient experimentation.

## PUBLICATIONS

---

**Palmbench: A comprehensive benchmark of compressed Large Language Models on mobile platforms**

*The Thirteenth International Conference on Learning Representations (ICLR), Singapore, 2025*

**Fake News Detection using a Deep Learning Transformer Based Encoder-Decoder architecture**

*IOS Press - Journal of Intelligent & Fuzzy Systems, 2023*

**End-to-End Speech Recognition of Tamil Language**

*Intelligent Automation & Soft Computing, 2021*

**Sign Language Translation using Multi Context Transformer.**

*20<sup>th</sup> Mexican International Conference on Artificial Intelligence (MICAI), Mexico City, 2021*

## EDUCATION

---

**University of Wisconsin-Madison**

**Sep 2023 – May 2025**

*Master of Science - Computer Sciences; CGPA: 3.827 / 4.0*

*Madison, WI*

*Courses:* Intro to AI, Advanced NLP, Foundational Models, Big Data, Data Cleaning & Integration for Data Science.

**SSN College of Engineering (Affiliated to Anna University), Chennai, India**

**Aug 2018 – Jun 2022**

*B.Tech in Information Technology; CGPA: 8.95 / 10.0*

*Chennai, India*

*Courses:* ML, Deep Learning, Probability, Statistics, Data Structures, Algorithms, Operating Systems and Compiler Design.

## CERTIFICATIONS

---

- Deep Learning Specialization — deeplearning.ai (Coursera)
- Natural Language Processing Specialization — deeplearning.ai (Coursera)
- TensorFlow Developer Specialization — deeplearning.ai (Coursera)
- Machine Learning – Stanford (Coursera)
- Python for Data Science – NPTEL / IIT Madras (Jan 2021)