

BADRI NARAYANAN MURALI KRISHNAN

 bmuralikrish@wisc.edu

 +1 848 - 363 - 8431

 mbadrinarayanan

 MBadriNarayanan

EXPERIENCE

Founding AI / ML Engineer <i>CuroNow</i>	Sep 2025 – Present	Madison, WI
• Architected a Medication Agent to provide reliable medication insights using LLM reasoning & fallback handling.		
• Built a centralized AI caching microservice that reduced duplicate LLM API costs by 40% and latency by 25% .		
AI / ML Engineer - Contract Role <i>BrainWaves Digital</i>	Jun 2025 – Sep 2025	Remote, USA
• Led the development of a Finance Agent from scratch to extract data & insights effectively.		
• Deployed a multi-model AI stack comprising Gemini 2.5 Pro/Flash, GPT-4 with intelligent fallback mechanisms.		
• Created a dashboard with ReactJS, Supabase SSO & Firebase to visualize the cash flow, balances, & transactions.		
Graduate Research Assistant <i>University of Wisconsin – Madison</i>	Sep 2024 – May 2025	Madison, WI
• Designed a scalable ingestion system for 80k+ medical documents, optimizing chunking, embedding, and retrieval.		
• Developed an Agentic RAG pipeline tailored for radiology reports, enabling clinicians to gain diagnostic insights.		
• Generated document embeddings and stored them in a Vector DB , serving locally hosted LLMs via Ollama .		
• Established an evaluation framework using NDCG@5 , achieving a score of 0.857 on 1K+ radiology reports.		
• Engineered an Orchestrator LLM to route queries across three specialized nodes (Retriever-only, RAG , and general-purpose LLM) for adaptive, clinician-friendly handling.		
Autonomous System Research Intern <i>Nokia Bell Labs</i>	Jun 2024 – Aug 2024	Murray Hill, NJ
• Constructed a multi-agent system leveraging autonomous agents for adaptive learning & collaborative reasoning.		
• Coordinated agent workflows with an Orchestrator LLM and Chain-of-thought prompting for multi-step planning and optimization. Agents executed specialized tasks via tool-calling interfaces		
• Built the framework with LangGraph & Streamlit to visualize agent interactions and analyze emergent behaviors.		
Associate Engineer – AI/ML <i>Qualcomm</i>	Jul 2022 – Jul 2023	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).		

TECHNICAL SKILLS

GenAI Competencies: GenAI, LLMs, RAG, Multi-Agent Systems, Prompt Engineering, Vector DB, ML

Programming Languages: Python, C, C++, TypeScript, ReactJS, PyTorch, Tensorflow

Frameworks & Tools: PyTorch, TensorFlow, LangChain, LangGraph, Spark, Git, Docker, Azure, AWS, GCP, CI/CD

Libraries: NumPy, pandas, matplotlib, scikit-learn, Streamlit, NLTK, spaCy, OpenCV

SELECTED PROJECTS

Relationship Extraction using Large Language Models Repo

- Extracting complex relationships between entities from unstructured text data, by fine-tuning **Llama3.2 LLM**.

Optimizing Natural Language Understanding: Fine-tuning Mistral 7B Repo

- Fine-tuning **Mistral 7B LLM** on the Samantha dataset for conversational and contextual question answering.

SELECTED PUBLICATIONS

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

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

Sign Language Translation using Multi Context Transformer.

20th Mexican International Conference on Artificial Intelligence (MICAI), Mexico City, 2021

EDUCATION

University of Wisconsin-Madison	Sep 2023 – May 2025	
<i>Master of Science - Computer Sciences;</i> CGPA: 3.827 / 4.0		Madison, WI
<i>Courses:</i> 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	Sep 2018 – May 2022	
<i>B.Tech in Information Technology;</i> CGPA: 8.95 / 10.0		Chennai, India