

# BADRI NARAYANAN MURALI KRISHNAN

 immbadri3@gmail.com

 +1 848 - 363 - 8431

 in/mbadrinarayanan

 MBadriNarayanan

## SUMMARY

AI/ML Engineer with an MS in CS and 2 YoE delivering production-grade GenAI solutions from 0 to 1. Experienced in **Multi-Agent orchestrations**, **RAG Pipelines**, **LLM fine-tuning**, **MCP** and **A2A protocol** with a track record of building reliable, scalable AI systems.

## INDUSTRY EXPERIENCE

<b>Founding AI / ML Engineer</b> <i>CuroNow</i>	<b>Sep 2025 – Present</b> <i>Madison, WI</i>
<ul style="list-style-type: none"><li>Architected a <b>Medication Agent</b> to provide reliable medication insights using <b>LLM reasoning</b> &amp; fallback handling.</li><li>Built a <b>centralized AI caching microservice</b> that reduced duplicate LLM API costs by <b>40%</b> and latency by <b>25%</b>.</li></ul>	
<b>AI / ML Engineer - Contract Role</b> <i>BrainWaves Digital</i>	<b>Jun 2025 – Sep 2025</b> <i>Remote, USA</i>
<ul style="list-style-type: none"><li>Led the development of a <b>Finance Agent</b> from <b>0 to 1</b> to extract data &amp; insights effectively from bank statements.</li><li>Engineered a <b>multi-model AI pipeline</b> (Gemini 2.5 + GPT-4) with intelligent fallback logic, improving extraction accuracy by <b>35%</b> and reducing manual review time by <b>44%</b>.</li><li>Created a dashboard with <b>ReactJS</b>, <b>Supabase SSO</b> &amp; <b>Firebase</b> to visualize the cash flow, balances, &amp; transactions.</li></ul>	
<b>Associate Engineer – AI/ML</b> <i>Qualcomm</i>	<b>Jul 2022 – Jul 2023</b> <i>Hyderabad, India</i>
<ul style="list-style-type: none"><li>Introduced evaluation frameworks and metrics for ML models to enhance efficiency and accuracy by <b>10.26%</b>.</li><li>Benchmarked <b>quantized</b> and <b>pruned models</b> on <b>Snapdragon SoCs</b> to assess latency, throughput, and power consumption under production workloads on SNPE (SnapDragon Neural Processing Engine).</li></ul>	

## INTERNSHIP AND RESEARCH EXPERIENCE

<b>Graduate Research Assistant</b> <i>University of Wisconsin – Madison</i>	<b>Sep 2024 – May 2025</b> <i>Madison, WI</i>
<ul style="list-style-type: none"><li>Designed a scalable ingestion system for <b>80k+</b> medical documents, optimizing chunking, embedding, and retrieval.</li><li>Developed an <b>Agentic RAG</b> pipeline tailored for radiology reports, enabling clinicians to gain diagnostic insights.</li><li>Generated document embeddings and stored them in a <b>Vector DB</b>, serving <b>locally hosted LLMs</b> via Ollama.</li></ul>	
<b>Autonomous System Research Intern</b> <i>Nokia Bell Labs</i>	<b>Jun 2024 – Aug 2024</b> <i>Murray Hill, NJ</i>
<ul style="list-style-type: none"><li>Constructed a <b>multi-agent system</b> leveraging <b>autonomous agents</b> for adaptive learning &amp; collaborative reasoning.</li><li>Coordinated agent workflows with an <b>Orchestrator LLM</b> and <b>Chain-of-thought</b> prompting for multi-step planning and optimization. Agents executed specialized tasks via <b>tool-calling interfaces</b>.</li><li>Built the framework with <b>LangGraph</b> &amp; <b>Streamlit</b> to visualize agent interactions and analyze emergent behaviors.</li></ul>	

## TECHNICAL SKILLS

**GenAI Competencies:** GenAI, LLM, LLM Fine-tuning, RAG, Multi-Agent Systems, Prompt Engineering, Vector DB  
**Programming Languages:** Python, C, C++, TypeScript, ReactJS, PyTorch, Tensorflow  
**Frameworks & Tools:** LangChain, LangGraph, Spark, Streamlit, Docker, Azure, AWS, GCP, CI/CD

## SELECTED PROJECTS

<b>Relationship Extraction using Large Language Models</b>	 <i>Repo</i>
<ul style="list-style-type: none"><li>Fine-tuned <b>Llama 3.2 (8B)</b> to extract structured entity-relation triples from unstructured text, improving baseline <b>F1-score</b> performance by <b>7%</b> through multi-epoch fine-tuning and benchmarking.</li></ul>	
<b>Optimizing Natural Language Understanding: Fine-tuning Mistral 7B</b>	 <i>Repo</i>
<ul style="list-style-type: none"><li>Fine-tuned the <b>Mistral 7B</b> model on the Samantha conversational dataset using a <b>QLoRA</b> pipeline with <b>4-bit quantization</b> and integrated <b>W&amp;B</b> for training metrics and checkpoints.</li></ul>	

## EDUCATION

<b>University of Wisconsin-Madison</b> <i>Master of Science - Computer Sciences;</i> CGPA: <b>3.827</b> / 4.0 <i>Courses:</i> Intro to AI, Advanced NLP, Foundational Models, Big Data, Data Cleaning & Integration for Data Science.	<b>Sep 2023 – May 2025</b> <i>Madison, WI</i>
<b>SSN College of Engineering (Affiliated to Anna University), Chennai, India</b> <i>B.Tech in Information Technology;</i> CGPA: <b>8.95</b> / 10.0	<b>Sep 2018 – May 2022</b> <i>Chennai, India</i>