**SRIRAM RAMPELLI**

Farmington, MI 48335 | Phone: 248-986-6414 | [LinkedIn](https://www.linkedin.com/in/sriram-rampelli-b41a75178/) | [GitHub](https://github.com/sriram7737) | [Email](mailto:sriramrampelli15@gmail.com) | [Portfolio](file:///C:\Users\srira\Downloads\sriram7737.github.io) | [HuggingFace](https://huggingface.co/sriram7737)

**PROFESSIONAL SUMMARY**

Experienced AI/ML Engineer specializing in generative AI, large language models, and scalable AI applications. Proven expertise in developing and optimizing Retrieval-Augmented Generation (RAG) models and fine-tuning LLMs for real-world applications. Demonstrated success in designing cost-efficient cloud solutions and leveraging innovative retrieval techniques—including vector databases—to enhance data-driven decision making. Adept in fast-paced, cross-functional environments with a strong foundation in Python, TensorFlow, and cloud deployment (AWS) and a keen interest in expanding expertise on Google Cloud Platform (BigQuery, Airflow).

**TECHNICAL SKILLS**

* **AI/ML & Frameworks:** PyTorch, TensorFlow, RAG Models, LLMs (GPT, Llama, BioBERT), (familiarity with Gemini AI)
* **NLP & Data Processing**: spaCy, NLTK, LlamaIndex, LangChain
* **Cloud & Infrastructure:** AWS; Familiarity with Google Cloud Platform (BigQuery, Airflow, Vertex AI), Docker, Kubernetes, Terraform, CI/CD Pipelines
* **Databases:** SQL, PostgreSQL; Familiarity with vector databases (Pinecone, FAISS, Weaviate)
* **Cost Optimization:** Cloud resource optimization strategies, serverless computing
* **Programming Languages:** Python, SQL

**EXPERIENCE**

**AI Intern – BCG GenAI Job Simulation | Feb 2025**

- Developed an AI-powered financial chatbot that automated SEC 10-K data extraction, reducing manual effort by 40%.

- Implemented NLP-based structured data extraction using spaCy, pandas.

- Automated financial sentiment analysisincreasing decision-making efficiency by **25%**, improving decision-making efficiency for financial teams.

-Explored retrieval-augmented solutions aligned with RAG principles to support dynamic query responses.

**Research Assistant – Lawrence Technological University | May 2024 – Dec 2024**

Developed an LLM-powered healthcare chatbot integrating fine-tuned LLaMA & enhanced BioBERT, improving medical Q&A accuracy by 23%.

Designed custom data pipelines to handle noisy clinical text, boosting OCR-based medicalcondition recognition by 35%.

Used bits and bytes, cuda to run the big LLM models to run on local machines

Published findings at IEEE CCWC 2025, focusing on LLM-driven healthcare automation.

Investigated innovative retrieval techniques, including vector-based methods, to enhance chatbot performance and published findings at IEEE CCWC 2025.

**Project Engineer – Wipro Technologies | Nov 2021 – Dec 2022**

- Developed an AI-based anomaly detection system, reducing system log errors by 35%.

- Optimized system performance by 27% through ML-driven monitoring framework.

- Managed 300+ test procedures & 100+ Agile sprint projects, ensuring high-quality releases.

**AI/ML PROJECTS**

**RSNA 2024 Lumbar Spine Classification (Kaggle)** | [GitHub](https://github.com/sriram7737/RSNA2024)

- Built CNN-based medical imaging model, achieving 92% accuracy on RSNA dataset.

- Improved model performance through ensemble learning & data augmentation.

**AI Chatbot** | [GitHub](https://github.com/sriram7737/Chatbot)

Leveraged the Natural Language Toolkit (NLTK), a pivotal tool in natural language processing, to develop an advanced chatbot.

This chatbot is designed to deliver detailed and informative responses, showcasing my ability

to apply complex NLP techniques to improve user interaction and automate customer service operations.

**Chatbot with GPT-2 & NLP** | [GitHub](https://github.com/sriram7737/Multi-Source-Data-Analytics-Chatbot)

- Developed an AI chatbot leveraging GPT-2 & Wikipedia data, enhancing document Q&A automation.

- Integrated speech-to-text & voice-based interaction for real-time AI responses.

**CERTIFICATIONS & PUBLICATIONS**

- [**Rampelli, S.** (2025). *Empowering Healthcare Data Systems with an Innovative Chatbot Application Utilizing Python and Advanced Generative AI Models.* Presented at **IEEE CCWC 2025**.](https://ieeexplore.ieee.org/document/10903815)

- NVIDIA-Certified Associate: Generative AI & LLMs

- Oracle Cloud Infrastructure 2024 Generative AI Certified Professional

**EDUCATION**

Master of Science in Computer Science | Lawrence Technological University | Dec 2024