SAMARTHA RAMKUMAR

206-227-3663 | samartha.arks@gmail.com

SKILLS

- 5+ years experience and training in Machine Learning for CV, NLP and Multimodal learning applications including image-to-text, speech-to-text, text-to-speech, LLMs, GAN, RAG, LoRA,
- Proficient in entity extraction, summarization, question answering, information retrieval, audio processing
- **AI frameworks** - Langchain and Haystack; Pytorch, Tensorflow, Hugging Face
- **Database: Vector**- Pinecone, Postgresql; **Graph** - Neo4j, Falkordb; **NoSQL** - MongoDB; **Caching** - Redis
- **Programming**: Python, Javascript/ Typescript, C, C++, Matlab
- **Cloud**: Microsoft Azure, Google Cloud Platform (GCP), AWS
- **Dev**: Scikit-learn, OpenCV, Spacy, NLTK, Streamlit, MySQL, Git, Linux, Bash, Docker, CI/CD, FastAPI

EDUCATION

University of Washington - *Seattle, WA* Sept 2021 - June 2023
Masters in ME – Data Science track GPA - 3.86/4.0
Relevant Courses: Deep Learning**,** Computer Vision, Natural Language Processing, Database Systems

BMS College of Engineering - *Bangalore, India* Aug 2016 – Sept 2020
Bachelors in ME – Machine Learning and Robotics track GPA - 9.04/10

WORK EXPERIENCE

USBANK (Contract) - *Dallas, TX*
Senior Consultant (Data Scientist) Mar 2025 - July 2025

- Fine-tuned small language models with SFT/RL for Agentic AI decision-making, improving domain specific task efficiency.
- Developed a graph-based, multi-modal RAG-anything pipeline/server, enabling scalable and flexible information retrieval across large-scale file systems and diverse unstructured data sources achieving more than 95% accuracy.
- Developed an in-house IAM platform (Identity and Access Management) to streamline onboarding for different internal teams onto the Agentic AI macro service, enhancing security and operational efficiency.

CITIBANK (Contract) - *Dallas, TX*
Senior Developer (Software Engineer - Data/ML) Nov 2023 - Mar 2025

- Leveraged OpenAI/ Gemini / Llama LLMs with Knowledge Graphs in Agentic AI workflows for developing agent assist and IVR platforms addressing personal banking card and transactions management.
- Implemented advanced LLM guardrails to enhance data integrity and compliance within the banking sector, resulting in a 30% reduction in compliance-related incidents.
- Realized 15% higher precision gains and relevance in model outputs for information retrieval from banking documents using graph based RAG and reranking techniques.
- Led foundational initiatives for the setup and deployment of the Citi LLM Gateway (R2D2), overseeing deployment, security and seamless API call handling in a high-availability enterprise setting.

NASA - JPL - *Seattle, WA*
Machine Learning Engineer - Industry Capstone Jan 2023 - June 2023

- Designed experiments on vision models combining knowledge distillation with semi-supervised algorithms like Contrastive Fixmatch, Mean Teacher, and masked DINO for deployment on the NASA-EELS robot with peak accuracy of 93%

SPORTSBOX AI, INC. - *Seattle, WA*
Machine Learning Engineer June 2022 - Sept 2022

- Built multi-modal 3D Golf pose generation pipelines with pose descriptions from LLMs and achieved 7% increase in accuracy
- ML Model tuning and hyper-parameter optimization; Data ETL; Unittests;

DELOITTE TOUCHE TOHMATSU INDIA LLP - *Bangalore, India*
Analyst Sept 2020 - Sept 2021

- Developed RESTful APIs using ML models for a Fortune 500 client for 3.5 million outlets across seven countries.
- Developed the item-item collaborative filtering recommendation logic in the Indonesian market with XGBoost, reducing the infrastructure cost by 20%, and achieving forecasting accuracy improvement by 5%
- Deployed and maintained ML pipelines on Microsoft Azure ML with Databricks (MLOps); Power BI

RESEARCH EXPERIENCE

Information Processing Lab - *Seattle, WA*
Graduate Research Assistant March 2022 - June 2023

- Designed experiments on monocular autonomous driving frameworks to perform 3D localization with depth estimation on road scenes using Transformer-based DETR-like models on the Kitti (+2.3% improvement over SOTA) and nuScenes datasets.

PROJECTS

Editable 3D dance pose generation on the multi-modal datasets

- Generated editable multimodal 3D Dance pose sequences combining LLMs and Denoising-Diffusion models on audio embeddings extracted from OPEN AI Jukebox (Beat Alignment Score - 0.27)

Customized Recipe Generator for healthy / junk foods deployed on the AWS Sagemaker with Streamlit API

- Used fine-tuned open-source LLMs like Llama 2 (BERTScore - 0.868) and Mistral models (BERTScore - 0.84)

Ensemble ALBERT on the Stanford Question-Answering Dataset (SQuAD 2.0)

- Boosted the performance of BERT/ALBERT models with Ensemble algorithms for the SQuAD Leaderboard (F1 score - 88.43)

US BANK SUMMARY

Identity service
Developed an in-house IAM platform to streamline onboarding across multiple internal teams onto the Agentic AI macro service. The platform automated user provisioning, enhanced security through role-based access controls, and improved operational efficiency by reducing manual processes. By integrating real-time audit logging and continuous monitoring, the system ensured secure and scalable access management, contributing to faster onboarding and reduced administrative overhead. Used mongodb for creating a locking mechanism for the scheduler to check for timely updates in approvals and permissions while running as multiple deployments.

RAG
Developed a graph-based, multi-modal RAG-anything pipeline/server to enhance information retrieval across large-scale file systems and diverse structured and unstructured data sources including URLs, texts, pdfs, excel, images, etc. The solution integrated advanced retrieval-augmented generation (RAG) techniques to support scalable and flexible querying, achieving over 95% accuracy in extracting relevant insights. By leveraging graph-based structures, the pipeline optimized data relationships and improved retrieval efficiency, enabling faster and more accurate access to critical information across a variety of sources.

Fine-tuned small language models using SFT)/RL techniques to enhance decision-making within Agentic AI workflows
Using Entity Extraction
In the Agentic AI workflow, fine-tuned models were used to extract key user information from raw inputs and make real-time decisions on task assignment. For example, when a customer provides input through a web form or chatbot, the model would automatically extract important details such as customer info, preferences, purchase history, and urgency level. This extracted data would then be passed along to the relevant internal systems for further processing.