Ganapathy Subramaniam Sundar

Portfolio Linkedin Github Medium

ganapathysubramaniam1999@gmail.com
+1 (437) 556-0264
Toronto

Checkout my recent Gen AI personal projects on Featured section of my linkedin and my website

PROFESSIONAL EXPERIENCE

AI/ML Development Intern | Nokia | Canada | January 2025 - Present

- **Scalable LLM Deployment:** Deployed Hugging Face's **Llama 3.1 405B instruct** model via **FastAPI** using **Transformers**; overcame resource constraints by optimizing the inference pipeline with **asynchronous processing** along with **8-bit quantization** and **KV cache reuse** reducing the latency per 100 tokens from 20 seconds to 8 seconds.
- **Integrated Langgraph AI Agent Ecosystem:** Built multi-tool AI agents with **LangGraph** for code execution, database queries, web scraping, and **dynamic report generation** with data visualization; resolved synchronization challenges by implementing **custom callbacks, error handling by maintaining task state checkpoint (state management).**
- **Built autonomous Crewai multi-Agent setup:** Engineered a **CrewAI**-based multi-agent framework for self-directed code execution; refined inter-agent messaging protocols to ensure seamless task coordination by designing a task manager agent. This multiagent setup boosted the process of rapid prototyping among the developers in our team by reducing the poc development period from **few days to few hours.**
- **Developed RAG based AI Assistant:** Created an in-house AI assistant using RAG architecture by internally hosting **ChromaDB** and using **LangChain** for implementing RAG chain; designed **custom vector schema** to retain domain specific knowledge, used **text-embedding-ada-002** embedding model to convert textual chunks into **dense vector embeddings** along with iteratively **optimized custom system prompts** to enhance context-aware user responses. This helped the business users to obtain **quick, accurate responses to domain-specific queries** to enhance their decision-making process.

AI/ML Developer | Credwise | Canada | Volunteering | August 2024 - November 2024

- Led a team of 6 AI engineers to develop an innovative RAG-based credit card cashback assistant (REWARDS AI) using **Gemini 1.5 pro** in **vertex ai**.
- **Automated web scraping** via **GCP jobs** to aggregate and normalize cashback and rewards data for Canadian credit cards, stored as JSON in GCP buckets.
- Deployed **vector search endpoint** on Vertex AI using **textembedding-gecko@003** for embedding ensuring precise context retrieval.
- Integrated real-time location and pricing data by developing GCP functions leveraging **Google Maps and Search APIs** for dynamic grocery recommendations.
- Solved data integration challenges by designing **custom batch processing using GCP cloud schedule**, overcoming inconsistencies across diverse data sources and APIs to avoid **inconsistent web page update** frequencies to yield real-time data (items, prices) and data field names **normalization by mapping** to avoid duplication, error handling by **Cloud Monitoring** and **Alerting, API retry mechanism.** *This ensured the users to get a smooth experience by faster query responses and reliable real time context.*

Sample Scenario: A user with CIBC and RBC cards querying "wheat bread and 2% milk" receives a personalized recommendation with nearby store details, pricing, and precise cashback information.

Python Engineer | Virtusa | Chennai, India | October 2021 - July 2023

- **Cloud Cost Forecasting**: Built a predictive time series model in AWS SageMaker to forecast cloud expenses using Grafana-sourced AWS usage data, pandas for preprocessing, and a **SARIMA model** with **AUTO-ARIMA tuning** and automated forecasting pipeline with lambda function for trigger. Resolved noisy data issues having inconsistent timestamps and irregular usage patterns due to on demand resource allocation by developing robust preprocessing and cleaning script to collect usage data into S3 buckets by using **Grafana's API** for usage data extraction and **pandas** for preprocessing and feature engineering. Achieved **RMSE** around **5** and **MAE** around **3.5**.
- **Cost-Effective ETL Framework:** Developed a Python-based data transformation **MVP** framework to migrate data from **AWS Redshift** to **Snowflake**, achieving a 170% reduction in ETL costs by eliminating license fee, minimized data transfer costs using **SQL Alchemy** and optimizing data flows using incremental ETL mechanism and automated with **AWS step functions**.
- **Advanced Data Security**: Engineered a data masking solution using **AWS Macie** by configuring Macie jobs to scan S3 data sources for PII information, implementing translation masking algorithm and **AES-256** encryption to secure sensitive information and ensure data compliance.

Junior Software Developer | Powertrac Engineers Pvt Ltd | IN, Tamil Nadu, Chennai | May 2021 - September 2021

- **Designed and Developed a Custom Database Access Application:** Involved in the creation of an in-house database access tool using Java's JDBC and MySQL, significantly streamlining data retrieval and management processes across enterprise systems.
- **Engineered Robust Software Solutions:** Developed and integrated scalable software solution using pandas and streamlit, focusing on enhancing functionality and features like data visualizations in streamlit UI.

Remote Organization Projects

Realtime AI News Anchor | Jan 2024- July 2024

- **Developed a Realtime AI News Anchor Platform:** Integrated with Reuters API for real-time news ingestion and leveraged **AWS Lambda** for serverless processing.
- **Engineered a Custom News Summarization Model: Fine-tuned the BART** transformer model from Hugging Face in **AWS Sagemaker** using **PEFT** and **PyTorch** to generate concise, news anchor-style summaries with training data stored in S3 Bucket.
- **Integrated Synthesia for AI Avatar and Video Generation:** Created visually engaging news presentations with a realistic female AI avatar.
- **Incorporated Eleven Labs for Natural Voice Narration:** Enhanced the platform with lifelike voice synthesis for a human-like news delivery experience.
- **Utilized AWS for End-to-End Deployment:** Implemented the entire platform within the AWS ecosystem, ensuring scalability and reliability.

EDUCATION

- **Postgraduate Degree,** ***Artificial Intelligence and Machine Learning*** **| Lambton College | CA, Toronto**
  (Aug 2023- April 2025) ***(Dean's Honor Student)***

- **Bachelor of Engineering (BE),** ***Computer Science*** **| Anna University| Chennai, India**
  (2016- 2020) **(Achieved 1st prize in inter-state level coding contest)**

SKILLS SUMMARY

Category	Skill Area	Specific Technologies & Expertise
Large Language Models (LLMs) & AI Technologies	**LLM Development & Deployment**	**Fine-Tuning Techniques:** RAFT, PEFT, Reward-based (Reasoning models), Traditional approaches
		RAG-Based Architecture
		Cloud-Based Deployment Endpoints: Vertex AI, AWS Sagemaker, Azure ML Studio
		Hyperparameter Optimization: Grid Search, Random Search, Bayesian Optimization
		Prompt Engineering: Chain of Thought, Few-Shot, Self-Consistency, Reason + Act (ReAct)
	LLM Frameworks & Architecture	**Open-Source Libraries**: Hugging Face Transformers, PyTorch, TensorFlow
		LLM Application Frameworks: LangChain
		Multi-Agent Systems: CrewAI, LangGraph
	Multimodal AI	**Image Generation:** Stability AI, OpenAI DALL·E 3
		Speech-to-Text: Faster-Whisper, OpenAI Whisper
		Text-to-Speech: OpenAI TTS-HD, Deepgram, Eleven Labs
		Translation: Google Translation API
		Voice AI Agent: OpenAI Realtime, Twilio
Data Science & Machine Learning	**Machine Learning Algorithms**	Regression, Classification, Clustering Algorithms using Scikit-Learn
	Time Series Forecasting	ARIMA, SARIMA, SARIMAX
	Natural Language Processing (NLP)	NLTK, TextBlob, VADER, Gensim, Textacy
	Data Reporting & Analytics	Power BI
Programming & Data Technologies	**Programming Languages**	**Primary:** Python
	Databases	**NoSQL:** MongoDB, Google Firestore
		Relational: SQLite3, MySQL
	Cloud Data Warehousing	Snowflake
Cloud Platforms & Integration	**Cloud Expertise**	**Google Cloud Platform (GCP):** Vertex AI Agents, Vector Search, Cloud Run functions and jobs, Storage Buckets, Google Firestore
		Amazon Web Services (AWS): AWS Bedrock, Sagemaker, S3 Buckets, Lambda, DynamoDB
		Microsoft Azure: Azure ML Studio, Model Catalog, Azure OpenAI Services
		Native AI Platforms: Google AI Studio, OpenAI Developer Console, Anthropic Developer Console, GitHub Models

CERTIFICATIONS (Obtained through Assessments)

- Data-OPS: Data kitchen – fundamentals, Data kitchen- data ops platform developer
- LinkedIn Skill Assessment: Python, Machine Learning, MySql
- Advanced python and problem solving: TestDome, Hackerrank (1) , Hackerrank (2) , StudySection