Ganapathy Subramaniam Sundar (AI Engineer)
Toronto, Canada | +1 (437) 556-0264 | ganapathysubramaniam1999@gmail.com
Personal website | LinkedIn | Github | Medium Articles Page

Professional Summary:

AI Engineer with 5+ years of experience across Financial, Telecommunications, Supply Chain, Pharmaceutical,and Electrical Logistics domains. Proven track record of fine-tuning large language models using PyTorch and TensorFlow, architecting RAG-based solutions with LangChain, and designing agentic workflows leveraging LangGraph and CrewAI. Skilled in orchestrating containerized microservices, implementing CI/CD-driven MLOps pipelines, real-time inference, edge AI deployment, hyperparameter optimization, and automated end-to-end ML lifecycle management to accelerate AI-driven innovation.

Skills Summary:

Programming Language and libraries	Python , Pandas , NumPy
Deep learning /AI Frameworks	PyTorch , TensorFlow , Keras , Hugging Face Transformers
Agentic AI Frameworks	LangChain , LangGraph , CrewAI , AutoGen , Pydantic , Langflow , Phi Data
RAG Architectures	Retrieval-augmented generation (RAG) , Agentic RAG , (CRAG)
LLM Finetuning techniques	PEFT (LoRA , QLoRA) , Reinforcement Learning from Human Feedback (RLHF)
Web/API frameworks	FastAPI
Prompt Engineering techniques	Chain of Thought, Few-Shot, Self-Consistency, Reason + Act (ReAct)
GCP	Vertex AI , AI Platform , Model Garden, Cloud Functions, Firestore
AWS	SageMaker , Bedrock , Lambda , DynamoDB , API Gateway , S3
Azure	Machine Learning , Functions , Openai Services
LLM APIs	Openai , Anthropic , Gemini , Perplexity , Grok (X) , Groq
Databases	MySQL , PostgreSQL , MongoDB , Cassandra
CI/CD , Version Control & Dev Tools	*GitHub Actions* , Git versioning , API development and integration
Data Reporting & Analytics	Power BI
Model Optimization & AI Safety	NVIDIA NeMo Guardrails , Quantization, pruning, distillation
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
GPU / Acceleration	NVIDIA CUDA / cuDNN

Employment History:

AI Engineer **JAN 2025 – JUN 2025**
Nokia, Canada

- **Scalable LLM Deployment:** Dockerized and containerized Hugging Face Llama 3.1 405B LLM instruct via FastAPI , PyTorch and Transformers; engineered an asynchronous, 8-bit quantized inference pipeline with KV-cache reuse to cut latency by 60% (20s → 8s per 100 tokens) with GPU acceleration using CUDA for improved performance.
- **LangGraph AI Agent Ecosystem:** Built full-stack, multi-tool agents for code execution, database orchestration, web scraping, and dynamic BI visualizations; implemented custom callbacks and Agentic Langgraph State management with human feedback.
- **CrewAI Multi-Agent Framework:** Engineered an autonomous, agentic workflow with CrewAI and a bespoke task-manager agent; optimized inter-agent messaging to accelerate PoC development cycles from days to hours.
- **LangGraph Orchestrated Agentic RAG-Powered ServiceNow AI Assistant:** Engineered an end-to-end agentic RAG solution in Azure ML Studio with LangChain and LangGraph, Azure Cosmos DB and custom vector schemas using text-embedding-ada-002; integrated ServiceNow REST APIs for dynamic ticketing, incident resolution, and workflow automation with high-precision (metrics BERTScore,BLEU), context-aware enterprise support.

AI Engineer **AUG 2024 – NOV 2024**
Credwise Financial, Toronto

- **Agentic RAG-Driven Rewards AI Leadership:** Led a team of 6 AI engineers to build an Agentic RAG-powered credit card cashback assistant (REWARDS AI) on Vertex AI with Gemini 1.5 Pro, driving personalized financial recommendations.
- **Automated Data Ingestion & MLOps:** Designed and orchestrated GCP-based microservice web-scraping pipelines to normalize and store Canadian credit-card rewards data as JSON in Cloud Storage using serverless cloud function; implemented CI/CD-aligned batch jobs via Cloud Scheduler for reliable updates.
- **Vector Database deployment**: Deployed a high-throughput vector search endpoint on Vertex AI using textembedding-gecko@003 embedding model.
- **Dynamic Contextualization & API Integration for agent tool calls:** Developed serverless GCP Functions integrating Google Maps and Places APIs for real-time location and pricing data, enhancing grocery spend recommendations.
- Cloud Monitoring and Alerting, Error handling and Troubleshooting by API retry mechanism

Machine Learning Engineer **OCT 2021 - JUL 2023**
Virtusa, India

- **Predictive Cloud Cost Forecasting:** Architected a scalable time-series forecasting pipeline on AWS SageMaker using Grafana-sourced usage metrics; leveraged serverless AWS Lambda with pandas for data ingestion and cleansing, applied SARIMA with Auto-ARIMA hyperparameter tuning, and automated forecast triggers via Lambda. Implemented robust preprocessing to handle noisy, irregular timestamps into S3, achieving RMSE ~5 and MAE ~3.5 for precise cost projections.
- **Cost-Optimized ETL Framework:** Engineered a production-grade, Python-based ETL solution to migrate data from Amazon Redshift to Snowflake, slashing ETL costs by 170% through license-free architecture. Utilized SQLAlchemy for optimized data flows, incremental load mechanisms for minimal transfer overhead, and AWS Step Functions for CI/CD-aligned, fault-tolerant orchestration.
- **Advanced ML-Driven Data Security & Governance:** Designed an enterprise PII detection and masking platform with AWS Glue jobs running custom NLP scripts, SageMaker based text clustering, and translation masking, secured by AES-256 encryption. Delivered automated data governance and compliance across S3 data lakes, ensuring sensitive information is identified, aggregated, and protected end-to-end.

Junior Data Scientist **FEB 2020 – SEPT 2021**
Powertrac Engineers Pvt. Ltd. , India

- **Transformer Health Prediction Using Time-Series Sensor Data:** Developed a predictive transformer maintenance model using XGBoost and SHAP for feature explainability on time-series sensor data,of temperature, voltage swings along with maintenance logs of repair dates and fault types stored in PostgreSQL database and Achieved ROC–AUC of around 0.80 in cross-validation for failure prediction reducing emergency repairs by 45% and unplanned downtime by 35%, with insights visualized via Power BI and presented to the stakeholders.
- **Anomaly Detection in Energy Consumption via Isolation Forests:** Built an anomaly detection system using Isolation Forest and One-Class SVM on historical smart meter data to flag potential energy theft and tampering, enabling proactive fraud detection in the power distribution network.

Organization Project:

- **Realtime AI News Anchor (Jan 2024–Jul 2024)**
  Developed a real-time AI News Anchor platform (Jan 2024–Jul 2024) that ingested live Reuters feeds via API and processed events serverlessly with AWS Lambda. I fine-tuned Hugging Face's BART transformer in AWS SageMaker using PEFT and PyTorch , storing training data in S3 to generate concise, anchor-style summaries. I integrated Synthesia to produce a realistic AI avatar for video delivery and Eleven Labs for lifelike voice narration, then orchestrated a fully serverless, end-to-end deployment within the AWS ecosystem to ensure high scalability, reliability, and low latency

---

Education:

PG Diploma in Artificial Intelligence and Machine Learning **AUG 2023- APR 2025**
Lambton College, Toronto, Ontario (**Dean's Honor Student)**

Bachelor of Engineering, Computer Science **JUN 2016 - MAY 2020**
Anna University, India – WES approved