GOUTHAM VIGNESH

gouthamvigneshs@gmail.com | 9952689236 | Bengaluru Karnataka

Summary

AI Research Engineer specialising in generative AI with 4 years of experience, I have gained extensive knowledge and skills in developing and deploying machine learning models to solve real-time business problems. My expertise includes building deep learning models, applying natural language processing techniques, and fine-tuning open-source large language models (LLMs) for various use cases. Hands-on experience in managing the entire project lifecycle of ML initiatives which involves collaborating with cross-functional teams, gathering requirements, designing experiments, implementing solutions, and deploying models in production-ready environments.

Experience

AI Research Engineer - Applied R&D Unit | Quantiphi - Bangalore, India | 01/2024 - Current

- Created Qnowledge Engineer (QE), an AI-driven agentic workflow tool that serves as an AI developer companion, facilitating the generation of production-ready applications from descriptions through skill planning, code generation, debugging, and container management, which helps other developers efficiently build prototype products for Quantiphi's **Codeaira** platform for internal use.
- Developed a multi-modal agentic workflow integrating video, image processing, multi-modal RAG, and layout-based information extraction with LLMs enhancing document understanding and Seamlessly integrated into internal platforms **QDocs** and **Dociphi**, enhancing their document processing and interpretation capabilities.
- Engaged in the development of both open-source and proprietary **multi-modal LLM** models to address diverse business use cases on Quantiphi's Generative AI platform, **Baioniq**.
- Developed an **Auto-Train** framework integrated with an auto-evaluation pipeline for multi-modal and text-based LLMs, enabling the generation of task-specific instructional data, automated evaluation of diverse datasets against respective metrics, and fine-tuning of customized language models.
- Developed and contributed to an internal **LLM Ops platform** designed to automate key aspects of model deployment, including **quantization, model merging, prompt optimization, and inference pipeline optimization**. This platform streamlines the deployment process by providing an automated framework for optimizing model performance and enabling seamless integration, helping users achieve efficient, production-ready solutions.
- Collaborated with domain experts and cross-functional teams to translate complex requirements into scalable AI solutions

Research Engineer (Gen AI) | SciSpace - Bangalore, India | 04/2023 - 01/2024

- Designed and executed projects/products employing large language models (open-source models or commercial APIs)
- Created a fine-tuning pipeline for **open-source Large Language Models** (LLMs) utilizing techniques like **PEFT** and **LORA**, achieving both enhanced task performance and a remarkable 75% reduction in training costs
- Developing scalable applications by integrating Large Language Models (LLMs) with the **LangChain** framework and connecting them to internal APIs
- Conducted experiments with **VLLM** and the **NVIDIA Triton Inference** Server to create an efficient LLM inference pipeline, resulting in a 30% reduction in response times and significant operational cost savings through streamlined model deployment
- Developing and deploying code libraries in Python, employing functions and classes, with a focus on AI-centric product development
- Leveraged cutting-edge AI tools and techniques to enhance model performance, staying up-to-date with the latest developments in the field.

Machine Learning Engineer | Vakilsearch - Chennai, India | 05/2020 - 03/2023

- Composed production-grade code to convert machine learning models into services and pipelines to be consumed at web scale
- Developed a **Semantic search engine** using sentence Transformers to retrieve relevant case documents to the user's query from the **Vector DB** based on the retrieve & re-rank pipeline using Bi-encoder and cross-encoder
- Applied **ensemble methods** that use different models (Random forest, SVM, Logistic regression) to classify different types of Indian address-proof documents based on text with 90% accuracy
- Created a Q&A chatbot with a **Retriever** and Reader pipeline
- Utilized Vector DB's Embedding Retriever to locate documents

- The **RobertaQA** model then analyzed retriever results, reducing backend agent workload by 60%
- Experienced building APIs in Python, particularly in **FastAPI** or Flask.

Machine Learning Intern | Vakilsearch - Chennai | 11/2019 - 05/2020

- Composed production-grade code to convert trained models into a serving pipeline which can be consumed at a web-scale
- Collaborated with the team on deploying machine learning models as Rest APIs using Fast API in production servers and in containerizing the applications using Docker.

Skills

- Python
- Rust
- Node JS
- Machine Learning Algorithms
- Deep Neural networks
- Transformers
- Large Language model
- Reinforcement Learning
- Pytorch / Tensorflow
- Scikit-Learn, Numpy, Pandas
- ML Flow, Kubeflow
- Pytorch lightning
- Langchain/ Langserve
- OpenCV
- Hugging Face
- Prometheus, Grafana
- Torch Serve
- Docker, Kubernetes
- Mongo DB
- Github
- Linux
- Fast API
- Vector DB
- Jupyter Notebooks
- Google Colab
- Gradio/ Streamlit

Education and Training

Sri Krishna College of Technology | Coimbatore, Tamil Nadu | 03/2020

Bachelor of Engineering: Computer Science

Publications

- **SciSpace Copilot: *Empowering Researchers through Intelligent Reading Assistance, AAAI Conference 2024.*** The tool accelerates research comprehension with advanced question-answering using the Retrieval Augmented Generation (RAG) approach, offering accurate answers, exact citations, and features like explanations, summaries, and annotation.
- **SciSpace Literature Review**: ***Harnessing AI for Effortless Scientific Discovery, ECIR 2024***, This introduces an advanced tool utilizing cutting-edge methods in vector-based search, reranking, and large language models to streamline the literature review process and facilitate efficient scientific discovery in the rapidly evolving academic landscape.
- **LSTM based approach for Generating Music from MIDI notes, IJAST, 2020 -** showcases the automatic generation of high-quality music using Recurrent Neural Networks trained on existing music data.

Accomplishments

- Received Best Performer award for the quarter ending in September 2021 while working on the Chatbot Development and Document Classification Project.
- Finalist in Smart India Hackathon 2019 for the idea to build a Virtual Medical Assistant which helps rural people with medical-related queries.
- Finalist in Hackference Data Science Hackathon conducted by Microsoft, Bangalore

Certifications

- IBM Applied AI Professional Certificate, Coursera, 07/2020
- Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning, DeepLearning.AI, 04/2020

Websites

- LinkedIn - https://www.linkedin.com/in/goutham-vignesh-243988128/
- Github - https://github.com/GouthamVicky