

Abhijith Roy

+61 435 600 851 | roy.abhijith11@outlook.com | linkedin.com/in/abhijith-roy | github.com/abhijith

EDUCATION

National Institute of Technology, Tiruchirappalli
Bachelor of Technology in Computer Science Engineering

Sept. 2019 - August 2023

AREAS OF INTEREST

Machine Learning	Natural Language Processing
Generative AI	Information Retrieval
Vector Databases	Statistical Models

TECHNICAL SKILLS

Languages: Python, C++, PostgreSQL, Cypher, JavaScript, R

Frameworks: Llamaindex, Autogen, Azure OpenAI, RESTful API, Langchain, OpenAI, FAISS, HuggingFace, Pytorch, TensorFlow, scikit-learn, pandas

Developer Tools: Neo4j, Docker, Spark, Jira, Git, Swagger, VS Code, MS Excel, SQLite, Pinecone

INDUSTRY EXPERIENCE

Deckle

AI Consultant

Melbourne, Australia

October 2025 - Present

- Built a RAG prototype leveraging VectorDBs and task embeddings to enhance LLM-driven task retrieval and generation.
- Developing a production-grade retrieval system enabling dynamic task orchestration from structured and unstructured client data.

JobSciences

Senior AI/ML Engineer

Melbourne, Australia

October 2025 - Present

- Engineered and deployed scalable indexing algorithms for skill-job title mapping as FastAPI microservices with precision@k and CTR tracking.
- Leading a team of interns to design ML lifecycles and taxonomies driving personalized job and skill recommendation pipelines.
- Prototyping a GPT-4-based job pathway predictor using Australian labor data and graph edge-removal algorithms for career mapping.

Carelon Global Solutions

Data Scientist

Bengaluru, India

July 2023 - September 2025

- Developed and productionized predictive models (ARIMA, XGBoost) using Python, PyTorch, and scikit-learn to forecast critical compliance and risk metrics.
- Automated complex claims and compliance workflows using AWS Textract, OCR, and systematic GPT-4o prompting to drive Document Intelligence.
- Engineered and deployed an end-to-end Generative AI Dashboard (UI/API) featuring Multi-Agent workflows orchestrated via Autogen and LangChain.
- Integrated graph database patterns (Neo4j) into the GenAI architecture and utilized QLoRA for RAG embedding fine-tuning for improved model grounding.
- Engineering and maintained vector data storage systems such as FAISS, Pinecone and SQLite to store vector embeddings for unstructured documents and spreadsheets.

PUBLICATION

- **Roy, Abhijith**; Deepak, Gerard, Santhanavijayan A (2023). Ontology Synthesis and Generation Using AI Orientated Hybrid Learning for Microeconomics, *Hybrid Intelligent Systems*
- **Roy, Abhijith**; Deepak, Gerard, Santhanavijayan A (2023). A Semantically Driven Model for Web Image Tagging Using Diverse Tag Selection, *NaBIC*

TECHNICAL PROJECTS

Diverse File Ingestion for FAISS VectorDB | Project

April 2025

- Developed a seamless pipeline that ingests any type of document to be stored in a vector database.
- Users can also ask questions and receive answers for the many number of documents uploaded.

Building Language Models from scratch. | Training

Dec 2023 - Jan 2024

- Developed different language models utilising Neural Networks, RNN, and Transformer-Decoder models.
- Language models were built using Pytorch framework and 5-gram embeddings as inputs.

Soft Prompting using GPT-2 model | Training

Dec 2023 - Jan 2024

- Implemented prompt tuning on a GPT-2 small model using PyTorch and fine-tuned it on three tasks: summarization, question answering, and machine translation.
- Evaluated the prompting on accuracy, validation scores and BLEU scores across the three different tasks.

CERTIFICATION

Advanced RAG Evaluation by DeepLearning.AI

Open Source LLMs with Hugging Face

Deep Learning.AI by Andrew NG (Coursera)

Mathematics for Machine Learning and Data Science

LLM Prompting for Developers by DeepLearning.AI