JOEL VARGHESE

Software Developer II

☎ +91 8078188537 @ joelvar541@gmail.com 🔗 linkedin.com/in/joel-eapen 🔗 https://leetcode.com/u/joevarghese

SUMMARY

Experienced **AI Engineer** specializing in **machine learning**, **deep learning**, and **LLMs** with contributions to **IBM Watson**. Skilled in optimizing AI systems for **real-time performance** and **scalability**. Expertise in **model optimization**, including **RAG fine-tuning**, **transfer learning**, and deploying **state-of-the-art architectures** like **transformers**. Proficient in **secure AI deployments**, delivering **production-level Python SDKs** to **data scientists and clients**, and leveraging **cloud platforms** such as **AWS** and **IBM Cloud**. Proven ability to design **scalable APIs**, enhance **AI model accuracy**, and deliver **low-latency solutions**.

EXPERIENCE

Software Developer II

IBM

01/2024 - Present India

- **Challenge**: Slow AI inference and inefficient similarity searches reduced real-time **performance by 30%**, impacting scalability and user experience.
  Solution: Deployed **Chroma DB** with vector indexing, improving inference speed by 40%. Integrated **OpenTelemetry** for observability and **MLflow** for experiment tracking, boosting system efficiency by 25%.
- **Challenge**: **LLM-generated outputs** lacked contextual accuracy, reducing effectiveness for **enterprise applications by 25%**.
  Solution: Developed optimized LLM prompts and option sets, improving contextual **relevance by 35%** and enabling **retrieval-augmented generation (RAG)** in **IBM Watson x.governance**.
- **Challenge**: **Security vulnerabilities** and **non-compliance risks threatened** AI model integrity and ethical use.
  Solution: Implemented secure **APPX packages** with encryption and integrated **IBM Watsonx AI Guardrails** for **bias detection**, **model protection**, and **real-time monitoring**, ensuring compliance with GDPR and HIPAA.

Software Developer I

IBM

01/2023 - 12/2023

- **Challenge: Model inference** times were too slow for production needs, impacting user experience.
  Solution: Implemented **distillation, quantization, and pruning**, increasing inference speed by 30%
- I**mproved real-time data handling** efficiency by **30%** for enterprise applications by optimizing backend systems.

EDUCATION

Bachelor Of Technology In Computer Science with Minor in VLSI and Embedded System

Apj Abdul Kalam Technological University

06/2019 - 06/2023 India

GPA
9.0 / 10.0

CERTIFICATION

IBM Machine Learning Specialist - Professional

Professional certification in Machine Learning

AI From the Data Center to the Edge – An Optimized Path Using Intel® Architecture

Course on Optimized Path Using Intel Architecture

KEY ACHIEVEMENTS

Achieved Knight level in LeetCode Contest, ranking in the top 4% globally

IBM Top Performer 2024

Achieved a **95% rating** for collaboration, resulting in a **20% increase in project efficiency**

Personalization and Customization of LLM Responses

Research paper published in the *International Journal of All Research Education & Scientific Methods*, and recognized as a top read on **ResearchGate**.

Image Classification of White Blood Cells Using Deep Learning Models

Research paper published in the *International Journal of All Research Education & Scientific Methods*, recognized for its **significant impact in the field.**

SKILLS

C++ | Python | Java | AWS | Big Data | Docker | SQL | Agile Methodologies | Deep learning | Machine Learning | PyTorch | Data Structures | Data Engineering | FastAPI

PROJECTS

AI-Driven Real-Time Inventory Management and Optimization System for Fulfillment Centers

Created an **advanced system for inventory management** and optimization.

- Developed a **real-time inventory management system using Large Language Models** for demand forecasting and product flow optimization, i**mproving forecasting accuracy by 20%.**

Offline Coding Assistant Utilizing oLLAMA in Visual Studio Code

Created an advanced coding assistant using the oLLaMA model.

- Developed an **Offline Coding Assistant integrated with Visual Studio Code** to deliver real-time **code suggestions, syntax corrections, and documentation references.**