Prem Kumar

Malaysia • premstroke95@gmail.com • +60183810069 • in/premstrk • premstroke.com

SUMMARY

Senior Machine Learning Engineer with **7 years** of expertise building AI models, developing and leading Generative AI solutions (**AI Chatbot & AI Agents**) for Enterprise and B2B companies (**United States, Malaysia & Singapore**). Seeking for remote **Senior & Lead Generative AI roles**.

EXPERIENCE

Senior Generative AI Engineer

Ancileo **February 2024 - Present, Singapore (Remote)**

- Developed automated travel insurance claim process using LangGraph AI Agent and RAG in Azure cluster.
- Optimised AI Agent to run 24/7 at $2/hour, reducing claims processing time by 60%.
- Applied advanced prompt engineering techniques to improve AI Agent response.
- Experienced in prompt engineering frameworks like LangChain, LlamaIndex and DSPy.
- Cross-evaluated open source LLMs on HuggingFace (Mixtral, Qwen, Llama).
- Experienced in integrating Celery and Async for distributed high-load processing.
- Deployed Agents on Kubernetes clusters with auto-scaling worker pods for scalability.
- Experienced in deploying models on Azure with CI/CD pipelines.

Generative AI Engineer

Azara AI **August 2022 - November 2023, United States (Remote)**

- Delivered Enterprise AI Agents with task-based workflows, enhancing operational efficiency.
- Pioneered the development of the first RAG pipeline for AI Chatbots with document retrieval and summarisation.
- Integrated AI agents with WhatsApp and Gmail for automated customer messaging service and increased engagement by 35%.
- Deployed LLM-Ops infrastructure on AWS using LangSmith, Grafana, Jaeger, and Prometheus for monitoring.
- Evaluated multiple LLMs using synthetic data with RAGAS, improving output quality by 35% through prompt optimisation.

Machine Learning Engineer

WISE AI **August 2019 - August 2022, Malaysia**

- Developed Liveness Detection for e-KYC. Contracted private and governmental clients based in Malaysia.
- Developed Facial Recognition models with DepthMap, 3DDFA and FaceNet.
- Lead the data collection process for model training and fine-tuning from over 30+ sources worldwide.
- Successfully achieved the Facial Recognition standard ISO 30107-3, becoming the first company in SEA.

Machine Learning Engineer

Neofin **July 2018 - August 2019, United States (Remote)**

- Implemented Facial Recognition solution for Loan Origination with a team of developers.
- Partnered with four notable US-based fintech and neo-banking institutions.
- Conducted research on machine learning algorithms, leading to a 20% increase in predictive model accuracy.

Graduate Researcher

Monash University **August 2017 - January 2018, Malaysia**

- Developed AI algorithm to perform multi-class Seed Classification using CNN to assist in seed grading system.

PROJECTS

Catering Chatbot

Caterspot (Singapore) • www.caterspot.sg/ • March 2024 - April 2024

- Developed Catering Chatbot to handle orders from customers. Core stack including LangChain and OpenAI LLM.

Support Chatbot

HostTempo (United States) • hosttempo.com/ • January 2024 - March 2024

- Developed AI Chatbot fine-tuned on Tempo platform support videos aimed to provide clear instructions to users.

SKILLS

Python, TensorFlow, FastAPI, Celery, AWS, Azure, Docker, OpenAI, PostgreSQL, Vector Database (Pinecone), JIRA, Git, Kubernetes, Redis

EDUCATION

Bachelor of Engineering - Computer Systems

Monash University • Malaysia • 2016