HARSHAN ASHWAD

Email: harshanashwad@gmail.com Mobile: +65 90480020 LinkedIn: harshan-ashwad-539629211/

CAREER SUMMARY

AI Engineer specializing in LLM fine-tuning, production-scale ML systems and edge deployment. Experienced in building ML pipelines from cloud training to on-device inference, optimizing model performance through systematic fine-tuning, and reducing operational costs through intelligent caching strategies. Expertise spans full-stack development, scalable AI architectures and end to end project delivery.

EDUCATION

National University of Singapore (NUS) **Jan 2024 – Jun 2025**
Master Of Computing in Artificial Intelligence

Thiagarajar College of Engineering (TCE) **Jul 2018 – Jun 2022**
Bachelor of Engineering in Computer Science

EXPERIENCE

AI Seer **Jul 2025 - Present**
Singapore
AI Engineer Lead

- Led backend development and enhanced Facticity (AI powered fact-checking platform), expanding verification reach by **300%** by integrating a Twitter bot, **increasing user engagement** from **~100 to 300+ daily interactions**
- Integrated AskLoky's market-intelligence API into the verification workflow, enabling crypto/token claim validation, **reducing false positives by 25%** and providing **evidence-backed** responses for **95% of crypto queries**
- Developed a Twitter-native helper bot serving FAQ-style queries, handling **150-200 fact-check requests daily**, and reducing dependency on **manual support by 60%**
- Optimized LLM orchestration layer by implementing **Redis/Typesense** caching for embeddings, reducing **OpenAI token costs by 40%** and improving **response time by 2.5×** under load
- Designed parallel search enrichment architecture in Facticity's pipeline with intent-aware retrieval, reducing query **processing time by 60%** and improving evidence diversity by **30%**
- Fine-tuned Llama 3.1 8B model on AWS SageMaker using 6,000 curated examples, improving classification accuracy from **57% to 75.83%** on Originality.AI benchmark
- Orchestrated **end-to-end model training** pipeline using **AWS SageMaker** on ml.g5.24xlarge instances, managing **dataset curation from MongoDB** and methodical experiment tracking, optimizing model for high precision (0.747) and perfect recall (1.000)
- Deployed fine-tuned LLM on Qualcomm Snapdragon X Elite edge device using GGUF quantization **(16GB→4.6GB)**, achieving **16 tokens/sec** on CPU with **43% performance gain through NPU acceleration** via Qualcomm AI Engine Direct SDK

IBM **Jan 2022 - Nov 2023**
Bangalore, India
Software Developer

- Managed Quote to Cash application using **Node.js and Vue.js**, taking full ownership after 4 months and developing **20+ RESTful APIs** integrated with IBM Sales Cloud backend services
- Created SMTP-based health monitoring system with Node.js cron jobs, **reducing downtime by 25%** through early anomaly detection and real-time webhook integrations
- Developed custom Python monitoring tool with ML-based anomaly detection (**Isolation Forest, ARIMA**) to track AWS resource usage, **achieving 20% reduction** in cloud costs through automated remedial actions
- Built interactive dashboard using **Dash and Plotly** to visualize cost trends and resource optimization impact, enabling data-driven infrastructure decisions
- Upgraded middleware and Vue.js rendering performance, **cutting bug resolution time by 30%** while managing AWS infrastructure (EC2, IAM, S3) for secure deployment

PROJECTS

ML Prediction Explainer (Explainable AI) [https://mlpeek.vercel.app/]

- Created an end-to-end ML prediction explanation pipeline enabling users to **upload tabular datasets** (up to 10,000 rows), **train models and receive SHAP explanations** via FastAPI, scikit-learn, React, and TypeScript
- Designed a multi-step UI/UX flow guiding users through dataset upload, target variable selection, model training, and explanation viewing, reducing model **debugging time by 40%** through **instant visual feature importance** breakdowns
- Implemented Random Forest classifier evaluating performance using accuracy, precision, recall, and F1-score, **achieving 92% F1-score on test datasets** with guaranteed 100% local accuracy for SHAP explanations

LLM-Powered Context Aware Q&A Assistant with Chat History (RAG)

- Architected a document-aware Q&A system using LangChain, LangGraph, and Hugging Face Transformers, reducing **average query latency to 3 seconds**
- Optimized domain-specific LLMs, achieving a **30% reduction in hallucination rates** and improved context relevance
- Reduced **query latency by 62%** through **vector caching** and hybrid search, increasing query success rate to 85%
- Scaled Q&A system to process 100 blog articles using **FAISS vector store** with 512-token chunking and overlap

Movie Profitability analysis (Data Science)

- Conducted in-depth analysis of TMDB 5000 Movie dataset across multiple dimensions (seasonal trends, genre-revenue relationships), **identifying summer revenue peaks** for Action/Adventure genres and informing data-driven budget allocation strategies
- Achieved 78.8% F1-score and 69.8% test accuracy predicting movie profitability using Random Forest classifier on 4,794 movies across 18 genres, **outperforming baseline Logistic Regression** (66.6% accuracy) by 3.2 percentage points
- Applied Association Rule Mining to **uncover genre co-occurrence patterns** (Romance+Comedy+Drama) and production company specializations, generating **strategic recommendations** for seasonal release optimization and genre pairing strategies

SKILLS

- **AI/ML Frameworks**: LLM finetuning (HuggingFace), openCV, Keras, PyTorch, Pandas, NumPy, scikit-learn, SciPy, Matplotlib
- **Retrieval-Augmented Generation (RAG):** LangGraph, LangChain, LlamaIndex, Vector databases (FAISS, Pinecone, Chroma, Typesense)
- **Databases**: MongoDB, DocumentDB, MySQL, SQLite
- **Development Frameworks:** Node.js, Vue.js, React, Flask, FastAPI, Docker, Microservices, HTML, CSS
- **Developer Tools & Monitoring**: Git, Sentry
- **Cloud & Infrastructure: AWS** (SageMaker, Bedrock, EC2, S3, IAM, ECS, AppRunner, CloudWatch, Cost Explorer), GCP, Azure, Redis, Typesense
- **Data Visualization**: Tableau and Power BI
- **Programming Languages**: Python, Javascript, TypeScript, C++, Java

CERTIFICATIONS

- **AWS** certified Cloud Practitioner (2022) [Validation number: 4G8VKNVKZ2B41F97]
- **Weights & Biases** Building LLM-Powered Applications (certificate 2025)
- **AlgoExpert** certificate for completing 100 coding questions (2022)

PROFESSIONAL ACTIVITIES

- Showcased trained Llama 3.1 8B model at Qualcomm AI Program for Innovators (QAIPI) APAC Demo Day, demonstrating on-device deployment with **cost and accuracy improvements** on Snapdragon X Elite
- Spearheaded the refinement of user story creation based on Agile learnings, resulting in a **25% reduction in backlog items** and a more focused approach to feature implementation