CHENG YUAN (ROSS) KING

Machine Learning Engineer | Data Scientist | NLP & Scalable ML

rosscyking@gmail.com +44 7467 348177 Sheffield, UK rosscyking1115 Ross King

PROFILE

MSc Artificial Intelligence candidate at the University of Sheffield with a Computer Science background and a strong interest in applied machine learning, NLP, and scalable AI systems. Experienced in building practical end-to-end ML workflows, from data preparation and feature engineering to model training, evaluation, and software integration. Recent work includes scalable PySpark pipelines, LLM-based event extraction, speech classification, document question answering, and recommendation systems. Seeking Machine Learning Engineer or Data Scientist roles where I can combine rigorous ML experimentation with reliable software engineering and real-world product impact.

PROJECTS

Scalable Machine Learning with PySpark

PySpark, Python, Stanage HPC, Slurm

2026

- Built distributed data mining and machine learning pipelines on datasets ranging from 1.9M to 20M records using Apache PySpark on the University of Sheffield HPC cluster
- Implemented web log mining, traffic prediction, HIGGS classification, and MovieLens recommendation tasks covering GLM, Logistic Regression, Random Forest, Gradient Boosted Trees, ALS, and k-means
- Designed robust evaluation workflows including temporal splits, stratified sampling, cross-validation, hyperparameter tuning, and test-set comparison
- Produced analytical outputs including access heatmaps, model performance comparisons, and user-cluster genre preference analysis

LLM-Based Event Extraction Baseline

Python, Qwen2.5-7B, Hugging Face, HPC

2026

- Developed a zero-shot LLM baseline for event extraction using Qwen2.5-7B-Instruct on MAVEN and WikiEvents datasets
- Compared unconstrained and constrained-label prompting strategies for trigger detection, event type prediction, and argument extraction
- Ran inference on NVIDIA A100 GPU resources through the University of Sheffield Stanage HPC environment using Hugging Face Transformers
- Built evaluation and error-analysis scripts to measure valid JSON rate, trigger accuracy, type accuracy, combined accuracy, and prediction failure patterns

Speech Speed and Tempo Classification

Python, scikit-learn, NumPy

2025

- Designed and benchmarked an end-to-end ML pipeline for speech classification using FBANK features and classical machine learning models
- Improved speed classification accuracy from 79.2% to 86.6% through feature standardisation, hyperparameter tuning, and model comparison

TECHNICAL SKILLS

Machine Learning: scikit-learn, PyTorch, model selection, evaluation, cross-validation, hyperparameter tuning

NLP & LLMs: Hugging Face Transformers, document QA, event extraction, prompt engineering, text processing

Data & Analysis: Python, pandas, NumPy, SQL, Jupyter, statistical analysis, data visualisation

Scalable Computing: PySpark, GPU computing, CUDA concepts, HPC workflows, Slurm

Audio & Signals: feature extraction, speech classification

Software Engineering: Java, JavaFX, Maven, Git, Docker, JUnit, Jest, Cypress

Languages: Python, Java, C++, JavaScript/TypeScript

LANGUAGES

English ●●●●●

Mandarin Chinese (Native) ●●●●●

Japanese (JLPT N1) ●●●●○

EDUCATION

MSc Artificial Intelligence

University of Sheffield

Sep 2025 – Sep 2026 Sheffield, UK

- Core modules: Scalable Machine Learning, Natural Language Processing, Parallel Computing with GPUs, Machine Learning, Data Science, Text Processing
- Focused on applied ML systems, NLP pipelines, scalable computing, and model evaluation using Python-based workflows

BSc Computer Science

Queen's University Belfast

Sep 2021 – Jun 2024 Belfast, UK

- Core modules: Data Structures and Algorithms, Software Engineering, Advanced Computer Architecture, Cloud Computing
- Built a strong foundation in object-oriented programming, algorithms, systems, software design, and security

Additional portfolio projects on GitHub include a document QA assistant, AI travel planner, and algorithm visualizer.

- Compared kNN, Logistic Regression, Linear SVM, and Random Forest models across accuracy and representation limitations
- Conducted error analysis showing that temporal mean pooling removed useful timing information, explaining weaker tempo-classification performance