Job Title: Machine Learning Engineer
Department: Text-to-Speech (TTS)
Location: Manchester, UK

About the Role:

As a Machine Learning Engineer, you'll help deploy and optimise LLM-based Text-to-Speech models that enable advanced voice cloning and natural-sounding speech synthesis. You'll work alongside like-minded researchers and engineers to bring new features to life, focusing on building scalable, production-ready systems that deliver seamless, low-latency voice experiences. You'll be instrumental in shaping the voice of the ConnexAI product used by our worldwide customer base.

Key Responsibilities:

- You will use tools like TensorRT, ONNX, TorchServe, and Triton to optimise and scale models for real-time, production-level deployment.
- You will implement and maintain CI/CD pipelines and deploy models on cloud infrastructure (AWS)
- You will monitor system performance, troubleshoot issues, and improve deployment strategies.

Requirements:

- A degree in Computer Science, Data Science, or a related field.
- 1+ year of deploying large-scale LLMs and/or TTS systems in production.
- Experience using Docker and Kubernetes (desirable)
- Experience with vLLM, TensorRT, ONNX, TorchServe, Triton, or similar tools.
- Python, cloud platforms, and performance optimisation experience

Why Join Us?

At ConnexAI, you will solve complex problems in a fast-growing, innovative industry. You'll directly impact the performance and scalability of our LLM-based TTS models, helping position us as a market leader. You'll be part of a collaborative, dynamic team that values your input and growth.

With the support, resources, and tools you need, you'll continuously evolve as a professional in the Machine Learning space, tackling new problems in the sector and advancing your career.