Zeren Shen

(519)-781-4782 | zeren71415@gmail.com | Toronto, ON | LinkedIn | GitHub

TECHNICAL SUMMARY

LLM & Machine Learning Engineer with expertise in deploying language models and optimizing low-latency systems. Experienced in building enterprise solutions, including voice assistants, RAG architectures, and LLM fine-tuning. Skilled in model optimization, cloud MLOps, and contributing to open-source LLM projects.

EDUCATION

University of Toronto *Sep 2021 - Nov 2022*

Master of Engineering, Mechanical and Industrial Engineering - AI/ML Focus

University of Waterloo *Sep 2016 - Apr 2021*

Bachelor of Mathematics, Statistics & Machine Learning, Minor in Computer Science

PROFESSIONAL EXPERIENCE

Machine Learning Engineer *May 2023 - Nov 2024*

ThinkGenAI Lab Inc. *Toronto, Canada*

Research Assistant RAG System

- Developed a RAG system that enables research teams to search, retrieve, and summarize newly published papers.
- Built an end-to-end pipeline for PDF ingestion, chunking, and vector embedding (Pinecone), integrating LangChain-powered ReAct agents for real-time retrieval, semantic search, and LLM-based Q&A.
- Containerized microservices with Docker, deployed a FastAPI backend for concurrent request handling and built a Discord bot with conversation summarization, memory management, and PDF export features.

LLM-Powered Dialogue Toy Robot

- Integrated Whisper-STT, GPT-3.5, and Google TTS for dialogue robot pipeline, achieving 1.2s end-to-end latency.
- Developed a high-performance FastAPI service with async/await to enable parallel voice processing and LLM inference, handling 20+ requests per minute through request batching.
- Reduced response time by 50% through optimized pipelining, leveraging audio chunk streaming and token-level generation for seamless user interaction.

Machine Learning Research Assistant *Sep 2021 - May 2022*

Laboratory for Extreme Mechanics & Additive Manufacturing, University of Toronto *Toronto, Canada*

- Implemented TensorFlow-based U-Net models for real-time X-ray image segmentation, achieving 0.93 MeanIoU in pore detection and reducing manual inspection time by 40% through automated defect-tracking algorithms.
- **Publication**: Zhang, J., Lyu, T., Hua, Y., **Shen, Z.**, *et al.* Image Segmentation for Defect Analysis in Laser Powder Bed Fusion. *Integr Mater Manuf Innov* **11**, 418–432 (2022). https://doi.org/10.1007/s40192-022-00272-5

PROJECT EXPERIENCE

Medical Domain LLM Fine-tuning *July 2024 - Jan 2025*

- Fine-tuned LLaMA-3-8B on PubMedQA dataset using QLoRA (4-bit quantization) and DeepSpeed (ZERO3), achieving scalable multi-GPU training on Lambda Cloud.
- Enhanced clinical QA accuracy by 18.7% and improved long-answer relevance by 24% (ROUGE) and 15% (BERTScore) via task-specific evaluation and optimization.
- Deployed with vLLM-Accelerated Inference and PagedAttention for low-latency medical text generation.

KEY SKILLS

- LLM Engineering: Model Serving (vLLM/TRT-LLM), RAG Architecture, PEFT/QLoRA
- MLOps: Git, Docker/Kubernetes, AWS SageMaker, MLflow, FastAPI, Microsoft Azure, CI/CD
- Frameworks: PyTorch, HuggingFace Transformers, LangChain/LangGraph/LangSmith