Jeng-Yue Liu

(412) 284-3635 | buffettl@andrew.cmu.edu | buffett0323.github.io | linkedin.com/in/buffettliu

Education

Carnegie Mellon University – School of Computer Science Pittsburgh, PA
Master of Science in Artificial Intelligence and Innovation May 2027
- GPA: 3.88/4.0; Selected Courses: Computer Systems, LLM Systems, Advanced NLP, ML/DL, AI Engineering, Gen AI, Diffusion

National Taiwan University Taipei, Taiwan
Bachelor of Business Administration in Information Management Jun 2025
- GPA: 3.82/4.0; Awards: Summa Cum Laude (**top 1%** of the school), Presidential Award, Bachelor Degree Thesis Award, Dean's List

Skills

Languages:	Python, C, C++, Java, JavaScript, TypeScript, R, Swift, SQL, Shell
Frameworks:	PyTorch, NumPy, Librosa, Hugging Face, LangChain, React, FastAPI, SGLang, JAX, vLLM, Diffusers, Flask
Infra/DevOps:	Docker, Kubernetes, Linux, Helm, Argo CD, Google Cloud Platform, GitHub Actions, Postman
Tools:	PostgreSQL, MySQL, Supabase, Qdrant, Apache Kafka, Prometheus, Grafana

Work Experience

Neutone Inc. Tokyo, Japan (Remote)
Research & Development Intern Dec 2025 – Apr 2026
- Ported the in-house real-time tone-morphing plugin to a SlowFast training pipeline, mitigating low-buffer granular artifacts and degraded timbre transfer to improve OOD reliability while preserving low-latency real-time inference.

Academia Sinica Taipei, Taiwan
Machine Learning Research Intern Jul 2024 – Aug 2025
- Outperformed state-of-the-art models on 2,500 hours of audio with a 47.3% reduction in multi-scale STFT loss, enabling controllable disentanglement in style transfer, by proposing a factorized codec with attribute-specific auxiliary tasks and information perturbation
- Achieved 86% k-NN top-1 similarity across 75k+ Beatport segments by designing a zero-shot timbre encoder with MoCo-v2 and Swin Transformer, leveraging sequence perturbation and temporal augmentations for timbre-invariant representation learning

Quid Inc. Taipei, Taiwan
Machine Learning Engineer Intern Dec 2024 – Jun 2025
- Reduced manual prompt tuning by 10+ hours per week by optimizing search result similarity ranking and match scoring with DSPy under Chain-of-Thought and MIPROv2, and automating summary and title generation through an LLM-based assessment module
- Improved emerging-hashtag Precision@50 by 18% by combining BOCPD change-point signals with creator-conditioned engagement features in a LightGBM classifier, enabling early detection of volatile trends at 1M+ TTCM creators

Project Experience

Frontier Constrained Decoding for Diffusion Language Models | DLLM, LLGuidance Feb 2026 – May 2026
- Developed a constrained decoding system for discrete diffusion LLMs using frontier masking and Viterbi-based joint span repair, achieving 95.3% schema validity on JSONSchemaBench (+18% over SOTA) with 3x faster inference
- Cut decoding latency 5.8× (mean) and 9.4× (p95) using async mask-GPU overlapping, AIMD multi-token unmasking, and zero-forward selective remasking for grammar violations over SOTA

Kernel Optimization for Sparse Attention on NVIDIA B200 | CUDA, Triton, LLM Inference Feb 2026 – May 2026
- Implemented a CUDA sparse attention kernel using WMMA tensor cores, cp.async double-buffered KV gathering, and split-K parallelization, achieving 22–50× speedup over naive pytorch implementation on NVIDIA B200
- Developed top-K indexer with FP8 dequantization and two-tier CUDA graph caching, reducing per-call dispatch overhead to near zero

GraphRAG for News Analysis with LLMs | LangChain, FastAPI, Next.js, Neo4j, Docker, FAISS Dec 2023 – Dec 2024
- Improved glossary adherence and cut token cost by 49%, achieving >70% expert-validated alignment, by fine-tuning LLMs with a glossary-first QA pipeline that retrieved glossary chunks and constrained answers to glossary definitions
- Eliminated 97% of manual analysis effort by engineering GraphRAG indexing and an LLM-powered full-stack app that extracted entities, distilled cross-article insights, and revealed shifts in public attitudes via temporal entity frequency analysis

Publications

- Guan-Ming Chiu, **Jeng-Yue Liu**, "Probing Functional Correctness in Diffusion Language Models". *ACL 2026 SRW* [*pdf.*]
- **Jeng-Yue Liu**, et al., "SynthCloner: Synthesizer-style Audio Transfer via Factorized Codec with ADSR Envelope Control". *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2026. [*arXiv*]
- **Jeng-Yue Liu**, Tzai-Hung Wen, "Trip-Purpose-Based Methods for Predicting Human Mobility's Next Location". [*B.S. Thesis*]