Chiung-Yi Tseng

ctseng@luxmuse.ai — https://ctseng777.github.io —
https://github.com/ctseng777 — https://huggingface.co/Chiung-Yi —
https://scholar.google.com/citations?user=LJpMtqIAAAAJ —
https://www.linkedin.com/in/ctseng612/

Research Interests

AI-assisted mathematical proof, autoformalization, interpretable reasoning, AI for science, LLM safety and alignment

Education

University of North Carolina at Chapel Hill
Master of Science in Computer Science — Sep 2009 – May 2011
National Central University
Bachelor of Electrical Engineering — Sep 2005 – May 2009

Publications

- Tseng, C., Thasin, M., Roy, S., Zhang, D., & Effiong, B. (2025). *StreetMath: Study of LLMs' Approximation Behaviors*. NeurIPS MathAI Workshop (poster). https://ctseng777.github.io/assets/pdf/StreetMath_2025.pdf
- Feuer, B., Tseng, C., **Lathe, A. S.**, Elachqar, O., & Dickerson, J. P. (2025). *When Judgment Becomes Noise: How Design Failures in LLM Judge Benchmarks Silently Undermine Validity*. arXiv preprint. https://arxiv.org/abs/2509.20293
- Tseng, C., Zhang, D., Bi, Z., & Song, J. (2025). *Diffusion-based Large Language Models Survey*. TechRxiv preprint. https://www.techrxiv.org/users/952417/articles/1321784
- Bi, Z., Chen, K., Tseng, C., Zhang, D., Wang, T., Luo, H., Chen, L., Huang, J., Guan, J., Hao, J., & Song, J. (2025). *Is GPT-OSS Good? A Comprehensive Evaluation of OpenAI's Latest Open-Source Models*. arXiv preprint. https://arxiv.org/abs/2508.12461
- Tseng, C., Song, J., Bi, Z., Wang, T., Liang, C. X., & Liu, M. (2025). *Active Learning Methods for Efficient Data Utilization and Model Performance Enhancement*. arXiv preprint. https://arxiv.org/abs/2504.16136

Professional Experience

Senior Principal Engineer — May 2024 – Present
SambaNova Systems, Palo Alto, CA

- Integrated multi-modal models (Llava-Med, Qwen2.5VL) onto SambaNova's custom AI accelerators. Conducted model pretraining, fine-tuning, evaluation, and deployment.
- Improved DeepSeek-V3 compiler fusion/tile strategies on aggregated continuous batching pipelines. Responsible for DeepSeek-V3 quality control: model monitoring, regression testing, debugging accuracy, and inference-speed degradation.

- Improved inference speed and debugging experience through compiler and tooling enhancements. Designed tracing tools for immediate visualization of intermediate computational graphs during the O0 development phase.

Staff Software Engineer Jul 2023 – Feb 2024
Stability AI, Remote

- Fine-tuned and deployed SDXL models with LoRA using ComfyUI and ONNX.
- Delivered NSFW filters and CSAM safety integrations.
- Represented the company at HackHer Hackathon; mentored participants.

Staff Software Engineer Nov 2019 – Nov 2022
Twilio, San Francisco, CA

- Led global infrastructure for Schrems II compliance.
- Collaborated on a BERT-based content optimization assistant; won an internal hackathon.

Software Engineer
Amazon Lab126, Santa Clara, CA May 2017 – Sep 2019
Amazon Web Services, Seattle, WA Jul 2016 – May 2017

- Built GeoIP/CDN rewriting systems to bypass the Great Firewall and improve customer experience.
- Developed multilingual ASR model gating criteria to ensure model quality.

Software Engineer / Senior Software Engineer Feb 2012 – Jul 2016
EMC, Cambridge, MA

- Led development of the Object Consistency Checker and self-repair algorithms.

Technical Skills

Languages: Python, Java, Lean, LaTeX
ML/AI Tools: PyTorch, HuggingFace, Triton, TensorRT, ONNX, vLLM, LangChain, CrewAI, Weights & Biases, lm-eval
Infrastructure: Docker, Kubernetes, Helm, Terraform, AWS, GCP, RunPod, OpenRouter, Slurm
Systems: REST APIs, Distributed Systems, DevOps, MLOps, CI/CD

Community Involvement

- Reviewer for AAAI 2026 and NeurIPS MathAI Workshop 2025.
- HackHer Hackathon mentor and volunteer.