Yifan Zhang (Skylar)

(217) 390-0877| yifanz28@illinois.edu | yifan-zhang-60b18b325 | skylarkie

Education

University of Illinois Urbana-Champaign (UIUC) Aug. 2024 - May 2026 (Expected)
Master in Electrical and Computer Engineering *Champaign, IL, US*
Courses: Applied Parallel Programming, Distributed Systems, Database Systems

Shanghai Jiao Tong University (SJTU) Sep. 2020 - Jun. 2024
B.S. in Computer Science and Artificial Intelligence *Shanghai, CN*
Machine Learning, Data Mining, Computer Vision, Natural Language Processing, Operating System, Data Structure

Skills

- **Languages**: Python, C/C++, Java, SQL, Shell, HTML, Golang
- **Tools**: AWS, PyTorch, CUDA, HuggingFace, Git, Docker, Linux, Slurm
- **Others**: Chemistry, Photograph, Pr/Ae, Guitar, DAW, Figma

Work

AWS, Amazon May 2025 –Aug. 2025
Software Development Engineer Intern - AI *Pheonix, AZ, US*

- Designed and deployed an LLM-based summarization system for Infrastructure-as-Code (IaC) analysis, generating digestible overviews of large-scale cloud deployments to assist region build.
- Integrated LLMs via AWS Lambda and Bedrock to extract insights from IaC logs and graph-based statistics, supporting both extractive and abstractive generation workflows.
- Developed a DynamoDB-based caching layer and RESTful APIs to enable low-latency, production-ready responses, supporting downstream UI integration.

Aviatrix Inc. Feb. 2025 –May 2025
Machine Learning Engineer Intern *Champaign, IL, US*

- Developed a deep learning-based anomaly detection pipeline for unstructured network logs, leveraging sequence modeling and semantic embedding to capture latent abnormal patterns.
- Evaluated the system on personalized real-world production logs from cloud controllers, demonstrating applicability to low-SNR environments.

Shanghai AI Laboratory Oct. 2023 –May 2024
LLM R&D Intern - OpenMMLab Team *Shanghai, CN*

- Led multiple evaluation studies in **OpenCompass (6.1k stars)**, focusing on robustness and alignment of frontier LLMs; improved accuracy on objective tasks and built unbiased batch inference for large-scale model comparison.
- Contributed to evaluation protocols for LLMs, assessing subjective robustness and functional capabilities (e.g., code generation) This assist upstream pre-training team with the development of **InternLM (7.1k stars)**.
- Proposed and authored the internal technical report on Circular-Eval, a method for improving multiple-choice robustness, later adopted in evaluation system.

Research

Organic Reaction Mechanism Elucidation with LLMs Apr. 2025 - Sep. 2025
Research Intern: Blender Lab @ UIUC | Supervisor: Heng Ji *Champaign, IL, US*

- Built the first large-scale, expert-curated dataset of organic reaction mechanisms (**oMe-Silver/Gold**), comprising 10k+ mechanistic steps with rich annotations.
- Proposed oMeS, a dynamic evaluation framework combining weighted Needleman–Wunsch alignment and Tanimoto similarity for partial credit, enabling fine-grained analysis of LLMs' mechanistic reasoning across 4 metrics.
- Analyzed performance of 10+ LLMs, revealing systematic failure patterns on domain-specific reaction reasoning. Conducted standard and COT fine-tuning on compact models, achieving consistent gains and outperforming some SOTA models under various conditions.
- Work submitted to **ICLR 2026** (under review). Contribute as the **1st author**.

Reaction-based Enzyme Sequence Generation and EC Prediction Ongoing
Research Intern: Blender Lab @ UIUC | Supervisor: Heng Ji *Champaign, IL, US*

- Proposed a novel **reaction-conditioned enzyme generation task**, enabling AI systems to generate enzyme sequences and corresponding EC numbers directly from chemical reactions.

- Constructed the first large-scale benchmark linking **chemical reactions, enzyme sequences, and EC numbers**, comprising 219k enzymes and 34k reactions curated from UniProt and Rhea.
- Benchmarked scientific LLMs (Text+Chem T5, Meditron-7B), revealing that domain-specific pretraining improves enzymatic reasoning but current models struggle on unseen reactions.

UnSE: Unsupervised Speech Enhancement Using Optimal Transport May 2022 – Oct. 2022

Research Assistance: X-LANCE Lab @ SJTU | Supervisor: Kai Yu & Wenbin Jiang *Shanghai, CN*

- Co-developed a **GAN-based** training method without paired training data according to optimal transport principle.
- Implemented this idea and conducted experiments on VoiceBank+DEMAND and prominent ASR benchmark.
- Conducted experiments that compared our proposed method with other GAN-based speech enhancement methods on the same benchmark.
- **Paper accepted by INTERSPEECH 2023**. Contributed as a 3rd author.

Iterative self-supervised learning for speech enhancement without clean speech Oct. 2022 – Jun. 2023

Research Assistance: X-LANCE Lab @ SJTU | Supervisor: Kai Yu & Wenbin Jiang *Shanghai, CN*

- Developed a self-supervised speech enhancement method that eliminates the need for clean targets, leveraging iterative refinement and uncorrelated noise mixing.
- Designed the complete training pipeline. Conducted experiments on real-world noise and a standard ASR benchmark. And achieved competitive scores against supervised baselines.
- **Contributed as 1st author** of the paper.

PROJECTS

Optimizing Prompts with LLM for Text-to-Image Generation

- Developed framework leveraging Multimodal Large Language Models (MLLMs) to optimize user prompts for text-to-image generation without additional training.
- Designed and implemented a pipeline for iterative prompt refinement, improving image aesthetics and alignment through scoring mechanisms such as CLIP and LAION Aesthetics models
- Conducted experiments on widely-used T2I models, analyzing the trade-offs between aesthetic quality and relevance across optimization steps. Presented findings to enhance usability of prompt engineering techniques in T2I models.

Exhibitopia: Exhibition Booth Management Platform

- Built a **full-stack** web application for managing anime exhibition reservations, supporting role-based permissions and real-world exhibition data integration. Developed exhibitor- and admin-facing interfaces for real-time inventory management, reservation control, and queue calling; implemented frontend in and backend in Typescripts.
- Designed and implemented complex database features: multi-condition triggers, stored procedures (e.g., atomic reservation handling), transactional integrity, and role-aware queue logic using MySQL.
- Deployed GCP with Docker and VPC, solving database access attacks through internal IP routing; implemented auto-translation.

WTP Protocol and P2P Music Player

- Implemented WTP, a reliable protocol with sliding window mechanism based on C++ UDP socket.
- Designed a P2P music player application based on WTP protocol, including UI design using Qt for Python and a distributed file system containing all music files on launched clients. Finally, all music in the database can be played synchronously on distributed clients.