Yiming Cheng

(+1) 773-690-7675 or (+86) 177-4086-7835 |eaminc0328@gmail.com |Chicago,Illinois

EDUCATION

Tsinghua University Department of Electronic Engineering ***Sep.2019-Jul.2024***

Bachelor of Engineering in **Electronic Engineering(Major)**

Minor in **Statistics ,** Minor in **Law *(Enrolled in School of Law; Transferred major to EE.)**

University of Chicago Department of Computer Science ***Expected:Jan.2026***

Master in **Computer Science(Pre-doc) Mlsys track**

- **GPA: 3.95/4.0**

PUBLICATIONS

- Kuntai Du,Bowen Wang,Chen Zhang ,**Yiming Cheng,..**Ion Stoica,Junchen Jiang,"LMPrefill: An Inference Engine for Prefill-only Workloads in Large Language Model Applications", accepted by SoSP 2025
- X Lan, **Y Cheng**, L Sheng, C Gao, Y Li "**Depression detection on social media with large language models**",arXiv preprint arXiv:2403.10750
- X Lan, J Piao ,**Y Cheng**,C Gao, Y Li et al. Recommendation for Inclusivity of Underrepresented Producers in Usergenerated Content Platforms. In recycle
- Yi Yang,Hao Feng,**Yiming Cheng**,Zhu Han "**Emotion-Aware Scene Adaptation: A Bandwidth-Efficient Approach for Generating Animated Short**", MDPI-sensors 2024
- Yi Yang, Yitong Ma,Hao Feng, **Yiming Cheng**, Zhu Han, "Minimizing Hallucinations and Communication Cost: Adversarial Debate and Voting Mechanisms in LLM-Based Multi-Agents," MDPI Applied Science 2025
- **Yiming Cheng**, " Research on Recommendation System Technology Based on Large Language Models," Graduation Design,Tsinghua University, 2024.
- **Patent:** Yi Yang, **Yiming Cheng,** Hao Feng, et al. "A Semantic Encoding and Decoding Framework for Converting Visual Content into Virtual Animated Visual Representations.

RESEARCH EXPERIENCES

Graduate Research

Lmcache Team ***Sep.2024-present***

Open Source Contributor and Research Assistant ***Advisor: Prof. Junchen Jiang***

Working on open-source project:

- **LMCache:** The first open-source Knowledge Delivery Network (KDN) that accelerates LLMapplications up to 8x faster, at 8x lower cost.
- **vLLM/production-stack:** Scale from single vLLM instance to distributed vLLM deployment without changing any application code.Now an official project under vLLM.
- Matain and contribute to Open-Source repo.Working on Router Scheduling and cloud deployment . Contributed 1262 lines of code.

Argonne National Laboratory ***May.2025-Oct.2025***

Research Assistant ***Advisor: Prof. Kexin Pei***

- **EnvGym:** Project Leader,build and design a efficient agent system for os-level environment setup
- 2025 Summer of Reproducibility (SoR) Fellowship under the Open Source Research Experience (OSRE) and REPETO programs,supported by United States National Science Foundation (individual award)
- to be update,poster is coming out soon

Undergraduate Research

Future Intelligent Lab(FIBLAB),Tsinghua University ***Jul.2022—Jun.2024***

Research Assistant ***Advisor: Prof. Yong Li***

Recommendation for Inclusivity of Underrepresented Producers in User-generated Content Platform

- Take the pioneering step to thinking of the inclusivity issue of underrepresented producers in **UGC** (user-generated content) platform.
- Propose to construct a heterogeneous graph that can enrich the relations of vulnerable populations, and further propose **graph neural networks(GNN)** to learn representations based on enriching features from

multi-hop neighbors.

City Socioeconomic Simulator based on Large Language Models

- Use UE to Build a visual model scene of Beijing (CBD district)
- Use python to write scripts for agents to interface with LLM and design the agents' memory mechanism to do POI recommendation.(POI means point of interest in the city)
- Design and plug in agent-based recommendation systems

Signal Processing Lab,Tsinghua University ***Mar.2022—Jun.2023***

Research Assistant ***Advisor: Prof. Yi Yang***

Emotion-Aware Scene Adaptation: A Bandwidth-Efficient Approach for Generating Animated Shorts

- Use the PyTorch framework, build an image element and emotion recognition model based on the CLIP model and InceptionV3,and use PAD (Pleasure-Arousal-Dominance) for emotion scoring.
- Enhance the generated semantics using the EmoCap model trained based on PAD scores for emotion style, ultimately achieving higher emotional coherence than the baseline on the received new video frames.

Wireless Networking, Signal Processing and Security Lab,University of Houston ***April.2022—Jun.2023***

Research Assistant ***Advisor: Prof. ZhuHan, NAS Fellow***

Scalable AI Generative Content for Vehicular Network Semantic Communication

- This project aims to establish a large-model-based semantic communication channel and test its accuracy on a vehicular dataset
- Build and test a channel in PyTorch that uses CLIP to convert original images into semantics and then uses Stable Diffusion to restore semantics back into images.

INTERNSHIP

Beijing Thunisoft Information Technology Co., Ltd.

Software Engineer ***July.2022—Sep.2022***

- Use Spring Batch to develop a batch job scheduling system supporting complex workflows and dependency management. Scheduled tasks are executed as planned using Cron expression triggers.
- Integrate Quartz scheduler for enhanced flexibility.
- Data integrity and stability are assured with Spring transaction management and JDBC operations.

Beijing SmartBow Information Technology Co., Ltd.

Software Engineer ***June.2023—Sep.2023***

- Refactor the Sunflower library(main functions include JSON parsing, MQTT, B-Stack device info parsing, and data transmission encryption) for the company's Internet of Things (IoT) data platform using Go-lang
- Perform functional and performance testing on the refactored Sunflower library.
- Collaborate with hardware interns to debug and ensure successful MQTT-based data transfer of bridge deflection, vibration frequency, and temperature data from LuZhou Bridge to the company's database.

[XXXX] [XXXX] (Remote)

Quant Devloper **May.*2025—Sep.2025***

- Replaced and optimized the company’s traditional linear regression and deep learning basend models with a gradient boosting framework(XGboost). Applied advanced feature extraction and model training on ultra-high-dimensional data (2000+ features), achieving a 21% performance improvement in the IC (Information Coefficient) metric.
- Integrated classical ARIMA estimation from financial econometrics to guide and optimize LSTM-based forecasting. On high-frequency time series data, the approach has yielded a 2% performance gain, with further model training and hyperparameter tuning in progress.

***This XXX is for NDA requirements,for now it is an anonymous quant company from China mainland**

OTHERS

Scholarship:

Merit-based Predoc Scholarship of $40,000 ,University of Chicago (2024)

United States National Science Foundation for SoR project (2025)

Academic Interest:

Previously as undergraduate: Data mining(Recommendation System,Emotion Awareness,EmbodiedCity)

Current and future: System for machine learning(distributed LLM deployment,distributed KV cache,efficient ml)

Machine learning for systems(machine learning for code generation and Operating System)

Programming Skills: Python(Pytorch,CuPy), Go(Docker,K8s),Git(Github action),Linux,C,C++,Matlab,Verilog etc.

Personal Website: https://eaminc.github.io/ includes github,google scholar and other detailed infomation