# TIANYAO SHI

Purdue University, West Lafayette • (765) 123-4567 • shi676@purdue.edu • linkedin.com/in/tianyao-shi-507967290

## RESEARCH INTEREST

My research interests span system, architecture, and sustainability. I am currently working on sustainable computing and LLM serving optimization, with a special focus on understanding and improving the energy- and carbon-efficiency of LLM inference and agentic LLM systems; building models to quantify the emerging perspectives of computing systems' environmental impacts such as biodiversity loss.

## EDUCATION

- **Purdue University, PhD Student in Electrical and Computer Engineering** 08/2024 – Now
  Advisor: Professor Yi Ding GPA: 4.0/4.0
- **Shanghai Jiao Tong University, Master in Electronic and Information Engineering** 09/2021 – 03/2024
  Advisor: Professor Xiaofeng Gao GPA: 3.69/4.0 Major GPA: 3.9/4.0
- **Shanghai Jiao Tong University, Bachelor in Computer Science** 09/2017 – 06/2021
  GPA: 88.73/100 (3.79/4.30) Major GPA: 89.65/100 (3.88/4.30)

## RESEARCH AND WORK EXPERIENCE

- **Performance, Energy, and Quality Considerations for Serving Quantized LLM** 03/2025 – Now
  *Graduate Research Assistant @ Purdue University, West Lafayette*
  - Developed qMeter, a fully automated online profiling framework to jointly evaluate the performance, energy efficiency, and output quality of LLM serving under dynamic workloads.
  - Profiled 11 quantization methods across 4 model sizes (7B–70B) on NVIDIA A100 and H100 GPUs.
  - Revealed critical workload and hardware tradeoffs, demonstrating that while activation quantization scales well with tensor parallelism, KV cache compression can compound tail latency overheads.
  - Formulated an energy-optimal resource allocation solver using Integer Linear Programming in Python with ~1k LOC, reducing cluster energy consumption by 6.3% compared to greedy heuristics while meeting strict SLOs.
- **Quantifying and Reducing Biodiversity Impact in Computing** 03/2025 – Now
  *Graduate Research Assistant @ Purdue University, West Lafayette*
  - Developed FABRIC, the first fab-grave LCA framework to model computing's biodiversity impact.
  - Formulated novel metrics linking chip manufacturing and AI workloads to measurable species loss.
  - Reduced biodiversity footprint by up to 85.5% via hardware optimization and workload scheduling.
- **Sustainable LLM Serving via Disaggregation** 08/2024 – 10/2025
  *Graduate Research Assistant @ Purdue University, West Lafayette*
  - Proposed a SLO-aware LLM serving framework, GreenLLM, to minimize carbon emissions by using old GPUs.
  - Wrote a profiler as a vLLM plug-in using RESTful APIs in 1.7k LOC using Python and Linux Shell.
  - Reduced the carbon footprint by up to 40.6% compared to the standard serving scheme.
- **Quant Factor Verification** 05/2024 – 07/2024
  *Quant Investment Engineer (Internship) @ Innoam Assets, Shanghai, China*
  - Trained classical and deep machine learning models on proprietary quant factors and verified their performance.
- **Enterprise Customers' IT Budgets Prediction for Cloud Services**, 08/2022 – 09/2023
  *Graduate Research Assistant @ Shanghai Jiao Tong University, Shanghai, China.*
  - Formalized a new problem to predict enterprise customers' IT budgets for public cloud services via observed consumption records to launch targeted campaigns at high-value customers.
  - Devised and implemented a two-stage framework, BSA-DaMaM, to address the coupling of high feature-missing ratio and heterogeneity in real-world data in 3k Python code.
- **QoS Prediction in Public Cloud** 01/2021 – 10/2022
  *Graduate Research Assistant @ Shanghai Jiao Tong University, Shanghai, China.*
  - Led a 10-men team, built and published Alioth-dataset through 400+ VM co-location experiments.
  - Designed Alioth, an open-source framework in Python that leverages explainable machine learning to estimate

application performance degradation and detect co-location interference in public clouds with 94.71% accuracy.

- **Collaborative Recommender Systems** 12/2019 – 01/2021
  *Undergraduate Student Researcher @ Shanghai Jiao Tong University, Shanghai, China.*
  - Contributed to the visualization and writing of two papers on SIGIR and DASFAA 2021.

## PUBLICATIONS

1. **Tianyao Shi***, Yanran Wu*, Sihang Liu, Yi Ding, Disaggregated Speculative Decoding for Carbon-Efficient LLM Serving, IEEE Computer Architecture Letters (**CAL**), Volume 24 Issue 2, pp. 369-372, 2025.
2. Yi Ding, Yanran Wu, **Tianyao Shi**, Inez Hua, Beyond Climate Change: A Holistic Framework for Evaluating the Environmental Impact of Computing Systems, The Magazine for Environmental Manager (**EM**), Air & Waste Management Association (A&WMA), December 2025.
3. **Tianyao Shi**, Ritbik Kumar, Inez Hua, Yi Ding, When Servers Meet Species: A Fab-to-Grave Lens on Computing's Biodiversity Impact, ACM SIGEnergy Energy Informatics Review (**EIR**), Volume 5 Issue 2, pp. 34-40, 2025.
4. Yi Ding, **Tianyao Shi**, Sustainable LLM Serving: Environmental Implications, Challenges, and Opportunities, IEEE 15th International Green and Sustainable Computing Conference (**IGSC**), pp. 37-38, 2024.
5. **Tianyao Shi**, Yingxuan Yang, Yunlong Cheng, Xiaofeng Gao, Zhen Fang, Yongqiang Yang, Alioth: A Machine Learning Based Interference-Aware Performance Monitor for Multi-Tenancy Applications in Public Cloud, the 37th IEEE International Parallel & Distributed Processing Symposium (**IPDPS**), pp. 908-917, 2023.
6. Xuehan Sun, **Tianyao Shi**, Xiaofeng Gao, Yanrong Kang, Guihai Chen, FORM: Following the Online Regularized Meta-Leader for Cold-Start Recommendation, The 44th International ACM **SIGIR** Conference on Research and Development in Information Retrieval, pp. 1177-1186, 2021.
7. Xuehan Sun, **Tianyao Shi**, Xiaofeng Gao, Xiang Li, Guihai Chen, GCAN: A Group-Wise Collaborative Adversarial Networks for Item Recommendation, International Conference on Database Systems for Advanced Applications (**DASFAA**), pp. 330-338, 2021.

## PREPRINTS

1. **Tianyao Shi**, Yanran Wu, Inez Hua, Yi Ding, Sustainability of Computing Systems: A Survey from Environmental Impact Perspectives, TechRxiv. February 11, 2026.
2. **Tianyao Shi**, Yi Ding, Systematic Characterization of LLM Quantization: A Performance, Energy, and Quality Perspective, arXiv preprint, arXiv:2508.16712.

## SKILLS

- Programming languages: C/C++, Python, Linux Shell, SQL
- Software: Pytorch, Scikit-learn, XGBoost, Matplotlib, Pandas, vLLM, TensorRT-LLM, NVML, LaTeX, Tikz, Git
- Soft: Adaptability, Leadership, Technical documenting, Academic presenting, Cross-department communication

## HONORS AND AWARDS

- Institute for a Sustainable Future Research Award of Purdue University 03/2025
- Excellent Graduate Student Scholarship of Shanghai Jiao Tong University (6/196) 11/2023
- Outstanding Graduate of Shanghai Jiao Tong University 06/2021
- Excellent Bachelor's Thesis of Shanghai Jiao Tong University 06/2021
- Second Prize of LCCUP'20 Team Coding Contest (Ranked 95/2575) 10/2020
- B-Class Excellent Scholarship of Shanghai Jiao Tong University (twice, top 10%) 11/2018, 11/2020
- Meritorious Winner of Mathematical Contest in Modeling (MCM) (top 9.7% of 26062 teams) 05/2018

## ACADEMIC SERVICES

- Reviewer of IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 12/2024
- External Reviewer of:
  - IEEE International Conference on Data Mining (ICDM) 10/2023
  - AAAI Conference on Artificial Intelligence (AAAI) 12/2021
  - IEEE Transactions on Network and Service Management (TNSM) 02/2021
  - IEEE Transactions on Network Science and Engineering (TNSE) 11/2020