# Hongpeng Jin

CS PhD Student at Florida International University
11200 SW 8th St, CASE 212D, Miami, FL 33199, USA
hjin008@fiu.edu +1 469-543-7960 hongpengjin.me


## Education

**Florida International University** **Aug. 2023 – Present**
*Ph.D. in Computer Science* *Miami, FL*

**The University of Texas at Dallas** **Aug. 2016 – May 2018**
*M.S. in Information Technology and Management* *Dallas, TX*

## Research Interest

- Large Language Models (LLMs)
- LLM-based Agent System
- Ensemble Learning
- Efficient Inference

## Research Experience

- **Efficient Deployment and Inference of Large Language Models**

  Research on optimizing the inference effectiveness, efficiency, and scalability of Large Language Models (LLMs) through strategies such as cloud-edge collaboration, distributed AI, and ML efficiency techniques (e.g., early exit).
  Supervisor: Dr. Yanzhao Wu

  Contributions:

  [1]**CE-CoLLM**: A LLM development method to optimize the inference efficiency and accuracy of LLMs on edge devices through cloud-edge collaboration and ML efficiency techniques, addressing diverse requirements such as inference accuracy, low latency, resource constraints, and privacy preservation.

  [2]**DA-MoE**: A dynamic expert allocation mechanism for Mixture-of-Experts (MoE) models that leverages attention-based token importance in Transformer architectures to dynamically adjust the number of experts per token, enhancing efficiency and predictive performance.

- **Ensemble Learning for Model Performance and Robustness**

  Research on improving key aspects of Ensemble Learning, including diversity measurement, ensemble selection strategies, voting mechanisms, training methodologies, overall ensemble performance, and applications.

  Supervisor: Dr. Yanzhao Wu

  Contributions:

  [1]**LREnsemble**: An ensemble construction framework, effectively utilizing diverse models, generated through learning rate (LR) tuning, to construct efficient and high-quality ensembles, avoiding the waste of sub-optimal trained models by leveraging their diversity for ensemble learning.

  [2]**Synergistic Diversity metric (SQ)**: An ensemble diversity metric, significantly improving ensemble accuracy and robustness to out-of-distribution samples by optimizing diversity among member models.

- **Hyperparameter Optimization Strategy for Model Training**

  Research on developing advanced hyperparameter optimization strategies to enhance the training efficiency, performance, and robustness of deep neural networks (DNNs), transformers, and large language models (LLMs).

  Supervisor: Dr. Yanzhao Wu

  Contributions:

  [1]**LRBench++**: A dynamic learning rate tuning framework, improving DNNs and LLMs training efficiency and achieving a balance between model accuracy and training cost.

## Publications

1. CE-CoLLM: Efficient and Adaptive Large Language Models Through Cloud-Edge Collaboration
   **Hongpeng Jin**, Yanzhao Wu.
   arXiv preprint arXiv:2411.02829, under submission
2. Efficient and Learning Rate Boosted Deep Ensembles
   **Hongpeng Jin**, Yanzhao Wu.

under submission

3. DA-MoE: Dynamic Expert Allocation for Mixture-of-Experts Models
   Maryam Akhavan Aghdam, **Hongpeng Jin**, Yanzhao Wu.
   arXiv preprint arXiv:2409.06669, under submission
4. Effective Diversity Optimizations for Deep Ensembles
   **Hongpeng Jin,** Maryam Akhavan Aghdam, Sai Nath Chowdary Medikonduru, Wenqi Wei, Xuyu Wang, Wenbin Zhang, Yanzhao Wu.
   2024 IEEE International Conference on Cognitive Machine Intelligence (CogMI 2024)
5. Rethinking Learning Rate Tuning in the Era of Large Language Models
   **Hongpeng Jin**, Wenqi Wei, Xuyu Wang, Wenbin Zhang, and Yanzhao Wu.
   2023 IEEE International Conference on Cognitive Machine Intelligence (CogMI 2023)

# Professional Services

- Reviewer: IEEE Transactions on Big Data, 2025
- Reviewer: ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2024
- Reviewer: Asian Conference on Machine Learning (ACML), 2024
- External Reviewer: The Web Conference (WWW), 2025
- External Reviewer: Association for the Advancement of Artificial Intelligence (AAAI), 2024
- External Reviewer: International Joint Conference on Artificial Intelligence (IJCAI), 2024
- External Reviewer: The Web Conference (WWW), 2024
- External Reviewer: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024
- External Reviewer: SIAM International Conference on Data Mining (SDM), 2024

# Awards

- IEEE TPS 2023 NSF Travel Award, November 2023

# Work Experience

**Cintra US.** **May 2022 – Aug. 2023**
*Data Scientist* *Austin, TX*

- Built machine learning models to improve the work efficiency of our business or operation team, including improving our dynamic pricing model, building incident detection models, developing analytics AI for auto-analysis reports, etc.
- Conducted statistical analyses (AB tests) to quantify driver behaviors and preferences and measure the impact of various external interventions

**HP Inc.** **Apr. 2020 – May 2022**
*Marketing Survey Data Analyst* *Vancouver, WA*

- Modeled large-scale email survey data to quantify the margin effect of each customer journey experience.
- Developed a response priority algorithm for customer reviews using supervised LDA topic modeling and statistical learning.

**Samsung Electronics America.** **Mar. 2019 – Mar. 2020**
*QA Engineer* *Plano, TX*

- Validated functions related to communication networks of Android devices across different wireless networks (GSM, WCDMA, 4G, and 5G) through software, field, and automation testing; analyzed emerging issues based on device logs and testing data to identify root causes.

**ZTE USA Inc.** **Apr. 2018 – Mar. 2019**
*Software Test Engineer, automation testing* *Richardson, TX*

- Developed and implemented an automation testing transition project (“AIO”) aimed at replacing manual testing with automated test cases, enhancing the efficiency and quality of the regression testing process.