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*

Nanjing Normal University — **Sep. 2012 – Jul. 2016**
B.A. in Tourism Management — *Nanjing, Jiangsu, China*

Research Interest

- Large Language Models
- LLM-based Agent System
- Ensemble Learning
- Efficient Inferencing

Research Experience

- **Efficient Deployment and Inferencing of Large Language Models**
  Research on optimizing both the inference processes and scalability of Large Language Models (LLMs) by leveraging strategies like cloud-edge collaboration, distributed AI, and ML efficiency techniques (e.g., early exit, quantization).
  Mentor: Dr. Yanzhao Wu
  Contributions:
  - **CE-CoLLM: Efficient and Adaptive Large Language Models Through Cloud-Edge Collaboration**, arXiv:2411.02829, under review at MLSys 2025
    Proposed the CE-CoLLM 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.
  - **DA-MoE: Dynamic Expert Allocation for Mixture-of-Experts Models**, arXiv:2409.06669, 2024
    Proposed the DA-MoE method, 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.
- **Advanced Training Strategies and Ensemble Learning for Model Performance**
  Research on enhancing training efficiency, performance, and robustness of deep neural networks (DNNs) and large language models (LLMs).
  Mentor: Dr. Yanzhao Wu
  Contributions:
  - **Efficient and Learning Rate Boosted Deep Ensembles**, under review at CVPR 2025
    Proposed the LREnsemble framework, effectively utilizing diverse models, generated through learning rate (LR) tuning, to construct efficient and high-quality ensembles, avoiding the waste of sub-optimal LR-tuned models by leveraging their diversity for ensemble learning.
  - **Effective Diversity Optimizations for Deep Ensembles**, CogMI 2024
    Proposed the Synergistic Diversity metric, significantly improving ensemble accuracy and robustness to out-of-distribution samples by optimizing diversity among member models.
  - **Rethinking Learning Rate Tuning in Large Language Models**, CogMI 2023
    Introduced the LRBench++, a dynamic learning rate tuning framework, improving DNNs and LLMs training efficiency and achieving a balance between model accuracy and training cost.

Research Activities

- 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

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, creating incident detection model, creating analytics AI model for auto-analysis reports, etc.
  1. Dynamic Pricing: built LightGBM (quantile) models to predict future demand and its confidence interval, enabling the identification of demand anomalies.
  2. Incident Detection: Developed an incident detection prediction system using real-time vehicle status data (acquired from Wejo), combined with incident history reports, and highway pavement data.
  3. Analytics AI: developed prediction models and explainability tools for business decision-making.
- Conducted statistical analyses (AB tests) to quantify driver behaviors and preferences, including peak-hour behavior and lane-changing patterns, while measuring the impact of various external interventions, such as large events and extreme weather.

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

- Modeled the large-scale email survey data to quantify the margin effect of each customer journey experience (including selection, purchase, setup, and usage) on NPS (Net Promoter Score) for providing teams or stakeholders suggestions about improvement direction.
- Sorted the response priority of customer review records by its metadata and the info extractions from its text data using a Supervised LDA topic modeling approach (a word embedding method) and statistical learning models for supporting the customer response team responding to customer complaints more effectively.
- Assist UX team in doing power analysis, conducting A/B testing on the survey's email title and UI, and measuring AA/AB testing results as the analytics specialist (HPS survey side) using general linear regression methods

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

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.

Awards

- IEEE TPS 2023 NSF Travel Award, November 2023