XUANSHENG WU

Email: wuxsmail@163.com **Phone**: +1 (213)-376-9286 **Homepage**: jacksonwuxs.github.io

EDUCATION

University of Georgia 08/2021 - 12/2025 (expect)

- *Ph.D. in **Computer Science*** (Advisor: Dr. *Ninghao Liu)*.

Shanghai University of International Business and Economics 09/2016 - 06/2020

- *B.S. in **Applied Statistics*** (Advisor**:** Dr. *Chengcheng Hao)*.

HONOR & AWARD

- Awarded **740+ citations** on Google Scholar 07/2025
- Awarded **870+ stars** on GitHub for first-author projects 07/2025
- Travel Award for attending KDD 2025 06/2025
- Best Poster Award of UGA SoC Day - **Top 4%** 04/2025
- Travel Award for attending WWW 2024 03/2024
- 2021 Tencent Advertising Algorithm Competition - **Top 2%** 07/2021
- Baidu Python Good Coder - **1/22** 03/2021
- Shanghai Aijian Scholarship - **Top 1%** 01/2020
- Research Pioneer Award of Shanghai University of International Business and Economics - **Top 1%** 08/2019
- 2nd Prize of China Programming Contest for College Students - **Top 4%** 08/2019
- Single-subject Scholarship of Shanghai University of International Business and Economics - **1/125** 11/2017
- 1st Asian University Archery Championship - **3rd** place in Men Group 04/2017

SELECTED PUBLICATION

Foundation Models: ***Mechanism Interpretation***[1, 2, 3, 4, 5, 6, 7,10, 13], ***Human Alignment***[1, 3, 4, 9, 11, 12, 13], and ***Evaluation***[6, 8, 9, 12].
Broader Interests: ***Recommender Systems***[11, 15, 20], ***Multi-Modality Modeling***[2, 8, 17, 19, 20], and ***Science Education***[6, 9, 16, 18].
Note: † indicates the first author publications, ‡ indicates the co-first author publications.

[1] †Extracting and Utilizing Interpretation in Large Language Models, **KDD PhD Consortium *(Oral)***, 2025.
[2] Concept-Centric Token Interpretation for Vector-Quantized Generative Models, **ICML**, 2025.
[3] †Self-Regularization with Sparse Autoencoders for LLM-based Classification using SAEs, **KDD *(Oral)***, 2025.
[4] †Interpreting and Steering LLMs with MI-based Explanations on Sparse Autoencoders, submitted to EMNLP, 2025.
[5] ‡Beyond Input Activations: Identifying Influential Latents by Gradient Sparse Autoencoders, submitted to EMNLP, 2025.
[6] †Unveiling Scoring Processes: Dissecting Differences between LLMs and Human Graders, **TKNL**, 2025.
[7] ‡A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models, ArXiv, 2025.
[8] A Large Multimodal Ophthalmology Dataset and Benchmark for Large Vision-Language Models, **NAACL Findings**, 2025.
[9] Artificial Intelligence Bias on English Language Learners in Automatic Scoring, **AIED *(Oral)***, 2025.
[10] †Understanding the Behavior Shift in LLMs after Instruction Tuning, **NAACL *(Oral)***, 2024.
[11] †Could Small Language Models Serve as Recommenders, **WWW *(Oral)***, 2024.
[12] InFoBench: Evaluating Instruction Following Ability in Large Language Models, **ACL Findings**, 2024.
[13] †Usable XAI: 10 Strategies Towards Exploiting Explainability in the LLM Era, submitted to Computing Survey, 2024.
[14] Retrieval-Augmented In-context Model Editing for Multi-hop Question Answering, **CIKM**, 2024.
[15] †DIRECT: Dual Interpretable Recommendation with Multi-aspects Word Attribution, **TIST**, 2024.
[16] Applying Large Language Models and Chain-of-Thought for Automatic Scoring, **CEAI**, 2024.
[17] Black-box Backdoor Defense via Zero-shot Image Purification, **NIPS**, 2023.
[18] †MeNSP: Matching Exemplar as Next Sentence Prediction, **AIED *(Oral)***, 2023.
[19] †A survey of graph prompting methods: techniques, applications, and challenges, ArXiv, 2023.
[20] †Rethinking the Impacts of Overfitting and Feature Quality on Small-scale Video Classification, **ACM MM *(Oral)***, 2021.
[21] †Lifestyle-based Approach for Cervical Cancer Screening, **ICDATA**, 2018.

INTERNSHIP

Amazon – Rufus Team Research Intern 05/2025 - 08/2025

- Developed a new mechanism interpretation method to understand how **reasonings** are performed within LLMs.
- Finding: Strong reasoning models (e.g., Qwen 7b) can better distinguish between "Strict", "Intuitive", and "Superficial" *logic rules* as humans by assigning different probability levels to different types of logic rules.

Samsung – Samsung Ads **Research Intern** 06/2024 - 08/2024

- Optimized a RAG system based on internal documentation of Samsung Ads by introducing query paraphrase, document reranking, and document semantic deduplication to improve retrieval quality, improved from 2.9/5.0 to **3.7/5.0**.
- Identify key challenges in developing benchmarks for real-world RAG systems: (1) limited human annotated domain datasets, (2) limited access to SOTA LLMs for privacy concerns. Evaluation is significantly suffered from hallucinations.

Tencent – Tencent AI Lab **Research Intern** 05/2023 - 08/2023

- Proposed a series of local and global explanation methods for interpreting transformer-based language models: (1) analyze the importance of each prompt token to the generated text, (2) understand the knowledge encoded by LLMs based on the activated patterns of the parameters of self-attention heads and feed-forward networks.
- Proposed to study the impacts of instruction tuning by comparing the explanations from the pre-trained and fine-tuned models. We found that: (1) instruction tuned models more effectively recognize instruction from user prompts to drive the response generation; (2) self-attention heads from shallow layers learn more word-pairs about instruction verbs than general verbs; (3) feedforward networks slightly rotate their pre-trained knowledge to adapt to downstream tasks.

Baidu - NLP **Applied Research Intern** 09/2020 - 04/2021

- Raised the Long Text Semantic Segmentation Task to alleviate the negative impact of distorted paragraphs on downstream searching services. Proposed a new training objective and a synthetic data construction strategy and achieved segmentation task **F1=72.31%**. This service is applied to support Baidu Top-1 Search and FAQ Mining System.
- Investigated the bottleneck of the low GPU utilization problem on the FAQ Mining System. Proposed a new architecture consisting of three components, namely Proxy, Scheduler, and Operator, which parallelly handles requests and schedules the GPU-based services dynamically according to the payload of each service. Designed a low-resource scheduling algorithm to assign services for multiple GPUs. Improved GPU Utilization by **2.9X** and QPS by **4.1X**.
- Worked with coworkers to improve searching experience of Baidu Top-1 Search under metrics **Good:Same:Bad=28:6:1** by optimizing services of Truncation Detection, Enumerate Answers, and Multi-Resources Recall.
- Deployed a query filter service for queries that require videos as answers (VQA) based on GBDT with **F1=93.43%**.

Pingan OneConnect Co., Ltd - Institute of Big Data **Applied Research Intern** 01/2019 - 06/2019

- Proposed using Feature Context-Free Grammar for the Seq2SQL task, new solution supports **34** external patterns.
- Completed intention recognition task with **Kappa=77.21%** through a feature-engineering pipeline.
- Developed an automatic machine learning module including correlation, clustering and time series analysis.

Shanghai Ronghao Investment and Management Co., Ltd **Data Analysis Intern** 01/2018 - 02/2018

- Designed and tested a short-term trading strategy by analyzing over **40** million millisecond transaction records.

SELECTED RESEARCH

Improving Generalizability of LLM-based Classifiers with Sparse Autoencoders (SAEs) 11/2024 - 02/2025

- Proposed to identify shortcut features in LLM embeddings with SAEs and regularize them in task-specific classifiers.
- Proposed a pre-train then fine-tune pipeline to ensure SAEs can capture task-specific features, and an auxiliary regularization term to remove the indirect impacts of these shortcut features toward classifier predictions.

Steering LLM Behaviors with MI-based Explanations on Sparse Autoencoders (SAEs) 06/2024 - 10/2024

- Performed theoretical analysis of SAE-learned features, and revealed that existing explanation methods for SAEs suffer from the frequency bias between semantical and lexical features so that semantical features fail to be interpreted.
- Proposed a mutual-information-guided objective to generate explanations for those semantical features.
- Improved LLM safety in jailbreak defense by activating interpreted semantical features with safety-related awareness.

Dissecting Differences between LLM Graders and Human Graders 02/2024 - 05/2024

- Evaluated whether LLM graders behave the same as human graders by comparing their crafted analytic rubrics.
- Observed that LLMs generally understand assessment items as humans, however, LLMs find shortcuts for automatic scoring if some human-graded samples are provided, highlighting the risks of using common LLMs for education.

Usable XAI 11/2023 - 02/2024

- Proposed the next step of explainable AI should be to let foundation models benefit from their explanations.
- Managed the entire team to conduct experiments of case studies to support our proposed utilities of XAI on LLMs.
- Key findings: (1) attribution scores between prompt-response tokens can be used to measure response quality (e.g., hallucination and correctness); (2) influence functions can be used to evaluate the generalizability of LLMs; (3) chain-of-thought strategy typically faithfully explain the behaviors of LLMs for end users.

Could Small Language Models Serve as System Cold-start Recommenders? 09/2022 - 09/2023

- Formalized the system cold-start recommendation problem and launched the first benchmark for this challenge.

- Provided a mathematical framework of the in-context recommendation under the Hidden Markov assumption and demonstrated that in-context recommendation ability could be enhanced by model and prompt pre-training.
- Proposed a corpus refinement method and a decoupled prompt pretraining method to enhance small language models, empowered BERT-mini (11M parameters) achieves comparable performance with BERT-large (330M parameters).

Matching Exemplars as Next Sentence Prediction for Automatic Scoring 04/2022 - 08/2022

- Proposed to develop scoring systems by using pre-trained language models to relax the need for training samples.
- Transformed the scoring task as multiple-choices problems achieved with Next-Sentence-Prediction task of BERT.

DIRECT: Dual Interpretable Recommendation with Multi-aspect Word Attribution 04/2022 - 08/2022

- Proposed a review-based interpretable recommender system, predicting user preferences by averaging sentiment polarities of words weighted by word importance, where a word is important if it corresponds to an aspect of the item.
- Employed a concept-bottleneck layer and maximized the coding rate reduction on the space of aspect representations by leveraging a word-word affinity graph extracted from a pre-trained language model to learn discriminative aspects.
- Quantified experiments and case studies showed that DIRECT is comparable to SOTAs but provides clear explanations.

NoPPA: Non-Parametric Pairwise Attention Random Walk Model for Sentence Representation 09/2021 - 11/2021

- Proposed a novel sentence encoder by integrating the non-parametric self-attention into the bag-of-words model.
- Applied a kernel function to construct contextual word embeddings based on both positional and semantic embeddings.
- Evaluated NoPPA on eight different tasks and exceeded the bag-of-words-based baselines by **Acc=2.89%** on average.

2021 Tencent Advertising Algorithm Competition: Multimodal Video Ads Tagging 05/2021 - 07/2021

- Investigated that the baseline InceptionV3+Vggish+BERT+Resnet50-->NextVLAD-->ContextGate-->MLP has two issues: overfitting and outdated feature extractor, used the following solutions to improve it from 76.10% to **GAP=82.10%**.
- Removed the center frame modal as a redundant modality; tuned dropout rate; applied data augment strategies.
- Upgraded feature extractor; Introduced Temporal Shift Model; Fused ASR captions within image flow.

SELECTED PROJECT

Reproduction of Forward-Forward Algorithm 01/2023

- Reproduced the Forward-Forward algorithm proposed by Geoffrey Hinton with Numpy, awarded **140**+ stars on Github.

DaPy: An easy-to-use data analysis framework for humans 09/2017 - Present

- Designed more user-friendly APIs than Pandas with Python built-in data structures, awarded **580**+ stars on Github.
- Achieved comparable efficiency with Pandas by using MemoryView, Cache-Friendly Operations, and Binary Search Index.

BeeDrive: Open Source Privacy File Transferring System for Teams and Individuals 09/2017 - Present

- Designed a lite-weight Python package for high-speed remote file management, awarded **10**+ stars on Github.
- Implemented a data encryption protocol based on MD5 and AES for file transferring, no TLS (SSL) certificate is needed.
- Implemented a built-in network address translation service to help individuals access their home-kept files from outside.

Distributed Archery Events Supporting System based on B/S+C/S Hybrid Architecture 12/2017 - 12/2018

- Proposed "B/S + C/S Hybrid Layout Architecture" to ensure high efficiency and high stability concurrently.
- Supported the 2018 Chinese University Archery Championships, handling peak system traffic of **1430** visits/second.

TEACHING & TALK

- Invited talk at research seminar of School of Computing, University of Georgia. 04/2024
- Teaching Assistant of CSCI 4380/6380 Data Mining, University of Georgia. 01/2022 - 05/2022
- Teaching Assistant of Archery Club, Shanghai Foreign Language Primary School. 09/2017 - 12/2018

SERVICE

- Transaction Reviewers: TNNLS, TKDD, TKDE, & TOIS.
- Conference Reviewers: ACL 2025, NAACL 2025, ICLR 2025, EMNLP 2024/2025, NIPS 2024, NIPS Workshops 2024, CIKM 2024, ICML Workshops 2024/2025, & ICLR Workshop 2024.

OTHERS

- Mandarin (fluent), English (fluent), Cantonese (novice).
- Python, Linux, SQL, C, SPSS, MATLAB.
- Archery (part-time coach 1+ years), Running (first half-marathon done), Basketball, Traveling w/ my girlfriend and family.