Zhuowei Chen

⋄ Email: zhuowei.chen@pitt.edu | ⋄ https://johnnychanv.github.io

EDUCATION

University of Pittsburgh Pittsburgh, PA
Ph.D. in Computer Science | Advisor: Prof. Xiang Lorraine Li *Sep 2025 - Current*

Guangdong University of Foreign Studies Guangzhou, China
B.E. in Software Engineering | Advisor: Prof. Lianxi Wang *Sep 2021 - Jun 2025*

University of California, Berkeley Berkeley, CA
Berkeley Visiting Student *Aug 2023 - Jan 2024*

PUBLICATIONS

* represents equal contributions and † represents the corresponding author.

1. **Zhuowei Chen**, Bowei Zhang, Nankai Lin, Tian Hou, Lianxi Wang.
   Unlocking LLM Safeguards for Low-Resource Languages via Reasoning and Alignment with Minimal Training Data.
   The Fifth Workshop on Multilingual Representation Learning, MRL Workshop @ EMNLP 2025.
   `Reinforcement Learning` `LLM Safety`
2. **Zhuowei Chen**, Qiannan Zhang, Shichao Pei.
   Injecting Universal Jailbreak Backdoors to LLMs in Minutes.
   The Thirteenth International Conference on Learning Representations, ICLR 2025.
   `Model Editing` `LLM Safety`
3. **Zhuowei Chen**, Yuben Wu, Xinfeng Liao, Yujia Tian, Lianxi Wang†.
   An Effective Deployment of Diffusion LM for Data Augmentation in Low-Resource Sentiment Classification.
   The 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024.
   `Generative Data Augmentation` `Diffusion LM`
4. Lianxi Wang, Yujia Tian*, **Zhuowei Chen***†.
   Enhancing Hindi Feature Representation Through Fusion of Dual-Script Word Embeddings.
   The Joint Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024.
   `MoE Representation Enhancement` `Low-Resource Languages`
5. **Zhuowei Chen**, Yujia Tian, Lianxi Wang†, Shengyi Jiang.
   A Distantly-Supervised Relation Extraction Method Based on Selective Gate and Noise Correction.
   The 22nd China National Conference on Computational Linguistics, CCL 2023.
   `Feature Selection` `Noise Reduction`
6. Xinfeng Liao, Xuanqi Chen, Lianxi Wang, Jiahuan Yang, **Zhuowei Chen**, Ziying Rong.
   OTESGN: Optimal Transport Enhanced Syntactic-Semantic Graph Networks for Aspect-Based Sentiment Analysis.
   The 25th IEEE International Conference on Data Mining, ICDM 2025
7. Lianxi Wang, Yujia Tian*, **Zhuowei Chen***, Mutong Li, Nankai Lin†.
   EditMDS: An Iterative Optimization Method for Multi-Document Summarization Based on Edit Operations.
   Data Intelligence.

EXPERIENCE

University of Pittsburgh Pittsburgh, PA
Graduate Student Researcher *Sep 2025 – Current*
Supervisor: Prof. Xiang Lorraine Li & Prof. Raquel Coelho

- Automatic Annotation Tool for Educational Peer-Feedback
  - Benchmarked six LLMs on annotation task across paradigms, including Zero-Shot, Few-Shot, Similarity RAG, PromptTuning, LoRA, Instruction Tuning, GRPO-based RL, and SFT–GRPO fused RL.

University of Massachusetts Boston Boston, MA
Research Intern *Mar 2024 – Oct 2024*
Supervisor: Prof. Shichao Pei

- JailbreakLLM: Exploring Novel Jailbreak Backdoor Attacks on LLMs.
  - Proposed a novel method to inject universal backdoors into LLMs without additional datasets or extensive computational overhead (lowest 5 samples with 30 seconds editing).
  - Executed comprehensive experiments, confirming a high jailbreak success rate (over 90% on Llama2-7b) and highlighting the urgency for advanced defensive strategies in LLMs.

AI Lab, Wisers Information Ltd. Hong Kong, China
NLP Research Intern *Dec 2023 - Mar 2024*

- Hong Kong Tourism Index Formulation
  - Built Roberta-based textual classification models with human-annotated social media content.
  - Applied transformers for time series regression to predict regional arrivals.

Guangzhou Key Laboratory of Multilingual Intelligent Processing Guangzhou, China
Undergraduate Research Student *Nov 2021 – Mar 2024*
Supervisor: Prof. Lianxi Wang

- Deploying Diffusion LM for Data Augmentation in Text Classification.
  - Fine-tuned LMs with a diffusion objective to capture in-domain knowledge and generate samples by reconstructing label-related tokens.
  - Designed attention-based mask schedule for the diffusion LM, balancing domain consistency, label consistency, and context diversity.
  - Conducted analyses and visualizations to study its underlying mechanism, followed by experiments validating its effectiveness across various low-resource scenarios.
- Enhancing Hindi Representations via Fusion of Pre-trained Language Models.
  - Proposed a method to enhance Hindi feature representation by combining Devanagari and Romanized Hindi pre-trained language models.
  - Ablations and extensive NLU task experiments show the superiority of our method, demonstrating the potential of multi-script integration to enhance low-resource language models.
- Distantly Supervised Relation Extraction (DSRE) with Learning-with-Noise Methods.
  - Combined selective gate and noise correction training framework for DSRE, which performs data selection and corrects noise labels during a three-stage training process.
  - Experiments demonstrated state-of-the-art performance, revealing a promising new approach for applying training-with-noise techniques in NLP.
- Multi-Label Text Classification (MLTC) with Knowledge Augmentation and Span Prediction.
  - Integrated span-prediction with an GNN-based knowledge augmentation module to enhance MLTC.
  - Conducted visualizations and analyses to study its working mechanism, emphasizing the critical role of incorporating domain-specific knowledge for LM.

SELECTED HONORS

- **Top Ten Outstanding Youth Award** Guangdong University of Foreign Studies, 2025
- **China National Scholarship** (Top 0.2%) Ministry of Education of the PRC, 2024

SERVICES

- **Reviewer.** ACL Rolling Review, ICLR 2026.