Zhuowei Chen

◇ Email: johnny.zhuowei.chen@gmail.com

EDUCATION

Guangdong University of Foreign Studies (GDUFS) Guangzhou, China
B.E. in Software Engineering. Advisor: Lianxi Wang *Sept 2021 - June 2025*
GPA: 3.80/4.00

University of California, Berkeley (UCB) Berkeley, CA
Courses: NLP, Introduction to AI, Computer Security *Aug 2023 - Jan 2024*
GPA: 4.00/4.00

PUBLICATIONS

* represents equal contributions and † represents the corresponding author.

1. **Zhuowei Chen**, Qiannan Zhang, Shichao Pei.
   Injecting Universal Jailbreak Backdoors to LLMs in Minutes.
   The Thirteenth International Conference on Learning Representations, ICLR 2025.
2. **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.
3. 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.
4. **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.
5. Lianxi Wang, Huayu Huang, **Zhuowei Chen†**.
   A Knowledge-Augmented and Label-Aware Framework for Multi-Label Text Classification.
   (Under review @ Journal of Computer Science and Technology)
6. Lianxi Wang, Yujia Tian*, **Zhuowei Chen***, Mutong Li, Nankai Lin†.
   EditMDS: An Iterative Optimization Method for Multi-Document Summarization Based on Edit Operations.
   (Under review @ ICASSP)

SELECTED HONORS

- **China National Scholarship** (Top 0.2%) Ministry of Education of the PRC, 2024
- **First-class Scholarship** (Top 4%) Guangdong University of Foreign Studies, 2024
- **First-class Scholarship** (Top 4%) Guangdong University of Foreign Studies, 2023
- **Silver Medal** (Top 5%) National College Computer Design Competition, 2022
- **Bronze Medal** China Undergraduate Mathematical Contest in Modeling (Regional), 2023

RESEARCH EXPERIENCE

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

- JailbreakLLM: Exploring Novel Jailbreak Backdoor Attacks on LLMs. (ICLR 2025)
  - Proposed a novel method to inject universal backdoors into LLMs without additional datasets or extensive computational overhead.
  - Developed a multi-node target estimation strategy that preserves attack stealthiness, effectively overwhelming and disabling the internal safety mechanisms of large language models.
  - Executed comprehensive experiments, confirming a high jailbreak success rate and highlighting the urgency for advanced defensive strategies in LLMs.

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

- Deploying Diffusion LM for Data Augmentation in Text Classification. (EMNLP 2024)
  - 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. (COLING 2024)
  - Proposed a method to enhance Hindi feature representation by combining Devanagari and Romanized Hindi pre-trained language models.
  - Conducted an in-depth comparison of different feature fusion techniques, including concatenation, summation, and cross-attention.
  - 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. (CCL 2023)
  - 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 adapted 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 PROJECTS

- BiasLLM: Adversarial Knowledge Editing Attacks on LLMs.
  - Combined GNNs with locate-then-edit techniques (ROME) to attack Llama-2, successfully exposing significant biases within the model.
- Multimodal NLP: Image-Text Interfacing with CLIP and Rational Speech Acts.
  - Used the CLIP model for image and caption retrieval, and further improved retrieval effectiveness by developing and applying a Rational Speech Acts inference procedure.

WORK EXPERIENCE

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

- Built BERT-based textual classification models with human-annotated social media content.
- Applied transformers for time series regression to predict regional arrivals.

LANGUAGES & SKILLS

- **Programming**: Python, Java, SQL, JS/HTML/CSS, C/C++, Golang.
- **Languages**: English(IELTS 7.5), Mandarin(Native), Cantonese(Native).

OTHER RELATED EXPERIENCE

- **Conference Attendance.** Poster and oral presentation on *EMNLP 2024* and *LREC-COLING 2024*.
- **Teaching Assistance.** TA for Language Processing Technique.
- **Member of Publicity at Student Union.** Organized poster presentations promoting AI equity.
- **Volunteer Lecturer at Dongguan Library.** Introduced basics of AI to the public.