Zhuowei Chen

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

RESEARCH INTERESTS

Safe & Trustworthy LLMs, Low-resource NLP, Data-efficient Learning, and Information Extraction.

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. Lianxi Wang, Yujia Tian*, **Zhuowei Chen*†**.
   Enhancing Hindi Feature Representation Through Fusion of Dual-Script Word Embeddings
   Proceedings of the 2024 Joint International Conference on Computational Linguistics.
   LREC-COLING 2024 (Long-paper, Main Conference)
2. **Zhuowei Chen**, Yujia Tian, Lianxi Wang†, Shengyi Jiang.
   A Distantly-Supervised Relation Extraction Method Based on Selective Gate and Noise Correction
   China National Conference on Chinese Computational Linguistics, 2023.
   CCL 2023 (Long-paper, Main Conference)
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
   (ARR June for EMNLP 2024, Avg. OA: 3.5)
4. Lianxi Wang, Huayu Huang, **Zhuowei Chen**†.
   LAKA: A Label-Aware and Knowledge-Augmented Framework for Multi-Label Text Classification
   (Under review)
5. Lianxi Wang, **Zhuowei Chen***, Yujia Tian*, Mutong Li, Nankai Lin†.
   EditMDS: An Iterative Optimization Method for Multi-Document Summarization Based on Edit Operations
   (Under review)

RESEARCH EXPERIENCE

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

- JailbreakLLM: Exploring Novel Jailbreak Backdoor Attacks on LLMs.

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

- BiasLLM: Adversarial Knowledge Editing Attacks on LLMs.
  - Combined GNNs with model editing techniques to attack Llama-2, successfully exposing significant biases within the model.
  - Highlighted the vulnerability of LLMs to adversarial knowledge editing, emphasizing the critical need for robust countermeasures.

- Deploying Diffusion LM for Data Augmentation in Text Classification. (ARR June)
  - 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 PROJECT

- 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 - Mar 2024*

- Quantization of Hong Kong Tourism Popularity.
  - Built BERT-based textual classification models with human-annotated social media content.
  - Applied transformers for time series regression to predict the number of regional arrivals.

SELECTED HONORS

- First-class Scholarship (Top 4%) GDUFS Academic Scholarship, 2023
- Silver Medal National College Student Mathematical Modelling Competition, 2023
- Silver Medal (Top 5%) National College Computer Design Competition, 2022

OTHER RELATED EXPERIENCE

- **Conference Attendance.** Poster and oral presentation on *LREC-COLING 2024* and *CCL 2023*.
- **Teaching Assistance.** TA for Language Processing Technique.

TECHNICAL SKILLS

- **Programming**: Python, Java, JS/HTML/CSS, C/C++, SQL, Golang.