

Zhuowei Chen

◇ Email: zhuowei.chen@pitt.edu | ◇ +1 (412)-589-7039

<https://johnnychanv.github.io>

EDUCATION

University of Pittsburgh

Ph.D. in Computer Science | Advisor: Prof. Xiang Lorraine Li

Pittsburgh, PA

Sep 2025 - Current

Guangdong University of Foreign Studies (GPA: 3.82)

B.E. in Software Engineering | Advisor: Prof. Lianxi Wang

Guangzhou, China

Sep 2021 - Jun 2025

University of California, Berkeley (GPA: 4.00)

Berkeley Visiting Student

Berkeley, CA

Aug 2023 - Jan 2024

PUBLICATIONS

* represents equal contributions and † represents the corresponding author.

- Zhuowei Chen**, Liwei Chen, Christian Schunn, Raquel Coelho, Xiang Lorraine Li.
[Neuron-Aware Active Few-Shot Learning for LLMs.](#)
The 64th Annual Meeting of the Association for Computational Linguistics, ACL 2026 (Oral).
Data Selection LLM Adaptation
- 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
- 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 (Main).
Generative Data Augmentation Diffusion LM
- 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 (Main).
MoE Representation Low-Resource NLP
- 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 Low-Resource NLP
- Zhuowei Chen**, Xiang Lorraine Li.
Neuron-Aware Data Selection for Annotation-Free LLM Self-Distillation.
The Forty-Third International Conference on Machine Learning, DEMO Workshop @ ICML 2026.
Data Selection LLM Adaptation
- Zhaoyi Joey Hou, **Zhuowei Chen**, Mengxue Zhang, Xiang Lorraine Li.
Rethinking Post-training Diversity Collapse: Is Diversity-preserving Post-training Enough?
The Forty-Third International Conference on Machine Learning, Human-AI Co-Creativity Workshop @ ICML 2026.
LLM Post-training LLM Creativity
- 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.
LLM Post-training LLM Safety

EXPERIENCE

Oracle America

Applied Science Intern | Supervisor: Dr. Chaithanya Bandi & Dr. Yingshen Wang

Seattle, WA
Sep 2025 – Current

- Benchmarking Self-Evolving OR Agents in Real Industrial Settings.

University of Pittsburgh

Graduate Student Researcher | Supervisor: Prof. Xiang Lorraine Li

Pittsburgh, PA
Sep 2025 – Current

- Low-Cost LLM Adaptation for Practical Educational Use.
 - Benchmarked six LLMs on annotation task across paradigms, including Zero-Shot, Few-Shot, Similarity RAG, Prompt Tuning, LoRA, Instruction Tuning, GRPO-based RL, and SFT-GRPO fused RL.
 - Developed a neuron-aware data selection framework for SFT that matched full-dataset accuracy with an 11% subset, demonstrating the utility of internal learning dynamics for data pruning.

University of Massachusetts Boston

Research Intern

Boston, MA
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.

Guangzhou Key Laboratory of Multilingual Intelligent Processing

Undergraduate Research Student

Guangzhou, China
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.

SELECTED HONORS

- **Top Ten Outstanding Youth Award** (Top 10/30000) 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, Data Intelligence.