Wenkai Li

wenkail@cs.cmu.edu | (412) 430-2334 | https://github.com/wenkai-li | https://wenkai-li.github.io

Research Interesting: NLP, LLM, LLM Reasoning, Responsible AI for Social Good, Multimodal Machine Learning

Education

Carnegie Mellon University, Language Technology Institution, School of Computer Science 08/2023 to recent

- GPA: 3.9/4.0
- Core Courses: Multimodal Machine Learning; Advanced Natural Language Processing; Introduction to Computer System; Large Language Model System; Neural Code Generation; Probabilistic Graphical Models

Northeastern University, Software Engineering 09/2019 to 06/2023

- GPA: 3.89/4.0
- College outstanding student scholarship (Awarded to the top 10% of students for academic excellence)
- Core Courses: Algorithm Analysis and Design; Discrete Mathematics; Data Mining Theories and Algorithms; Natural Language Processing; Web Development Programming Practice

Publications

- Jiarui Liu*, Iman Ouzzani*, **Wenkai Li***, Lechen Zhang, Tianyue Ou, Houda Bouamor, Zhijing Jin, Mona Diab. "Towards Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset" under review in 2025 ACL.
- **Wenkai Li,** Liwen Sun, Zhenxiang Guan, Xuhui Zhou, Maarten Sap. "1-2-3 Check: Enhancing Contextual Privacy in LLM via Multi-Agent Reasoning" under review in 2025 ACL.
- **Wenkai Li*,** Jiarui Liu*, Andy Liu, Xuhui Zhou, Mona Diab, Maarten Sap. "BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data" under review in 2025 ACL.
- Jiarui Liu, **Wenkai Li**, Zhijing Jin, Mona Diab. "Automatic Generation of Model and Data Cards: A Step Towards Responsible AI" published in 2024 NAACL.
- Hongliang Chen*, **Wenkai Li***, Leyi Zhang*. "Deep methods based on GAN for face-spoofing" published in 2022 International Conference on Machine Learning and Artificial Intelligence
- Guoqiang Liu*, **Wenkai Li***, Ruochen Xiao*, Youcheng Zhang*. "A New Approach for Text Style Transfer Based on Deletion and Generation" published in 2023 Computational Linguistics and Natural Language Processing

Research Experience

Large Language Model Negotiation with Diplomacy Setting Project 07/2024 to recent

*Research Assistant, **Language Technology Institute**, Carnegie Mellon University. Advisor: Daniel Fried, Mona Diab*

- Integrated the Sotopia framework with the CICERO model in the diplomacy setting, developed a pipeline to evaluate LLM dialogue quality by calculating diplomacy policy scores derived from dialogue.
- Implemented two-agents multi-turns negotiation pipeline and AI-human interaction framework.
- Building a negotiation evaluation benchmark for LLM, focusing on strategies, reasoning, and friendliness.

Make Large Language Model Keep Secret Project 08/2024 to recent

*Research Assistant, **Language Technology Institute**, Carnegie Mellon University. Advisor: Maarten Sap*

- Developed Extractor, Executor, and Checker agents for contextual privacy enforcement, reducing leakage by 36%.
- Implemented pipeline contains event extraction, privacy classification, and summarization for LLM privacy preserving.
- Conducted ablation study on information flow between the three agents and iteratively optimized the pipeline.

Large Language Model Personality Project 06/2024 to 10/2024

*Research Assistant, **Language Technology Institute**, Carnegie Mellon University. Advisor: Mona Diab, Maarten Sap*

- Led the project and addressed the challenge of realistically human personality simulation and reasoning correlation.
- Developed a large-scale dataset of 100,000 dialogues that reflect a wide spectrum of personality expressions to ground LLMs in realistic human personality expression through supervised fine-tuning and direct preference optimization.
- Demonstrated that training-based personality alignment methods outperform prompt-based approaches in assessments such as BFI and IPIP-NEO, with findings highlighting trait-based impacts on reasoning tasks.

- Demonstrated that ground with prompt-based and training-based alignment methods will affect the reasoning capability.

Model Card Generation and Translation Project 08/2023 to 12/2023

*Research Assistant, R3Lab, **Language Technology Institute**, Carnegie Mellon University. Advisor: Mona Diab*

- Spearheaded the design of the CardGen pipeline, integrating Large Language Models and Retrieval-Augmented Generation for automated model/data card creation.
- Constructed a dataset of 10,000 model cards with direct links to corresponding papers and GitHubs.
- Developed a hierarchical retrieve-and-generate system to automatically generate model and dataset cards.
- Evaluated the proposed pipeline using standard faithfulness metrics, GPT-based metrics, and human evaluation, demonstrating its effectiveness and comprehensiveness.
- Collected AI terminologies at scale and translated them into Arabic, Chinese, French, Japanese, and Russian through a combination of LLM-based and human validation, exploring its integration and applications in machine translation.

Text Style Transfer Analysis Project 06/2022 to 10/2022

*Research Assistant, **NLP Group**, Massachusetts Institute of Technology, Advisor: Gary Becigneul*

- Conceived a novel two-phase style transfer model, adeptly transforming text from Mark Twain's style to Wikipedia's, segregating the task into distinctive deletion and generation processes for enhanced linguistic fidelity.
- Integrated a Sequence Classification model with SHAP during the deletion phase to discern and excise style-indicative elements, setting a foundation for style-accurate content generation.
- Fine-tuned BERT’s Masked Language Model for stylistic consistency and leveraged SpaCy and GloVe for semantic coherence in the generative phase.
- Achieved a 10% improvement in style conversion efficiency over current state-of-the-art models, and refined performance through grid search optimization while preserving semantic integrity and readability.

Multimodal sentiment analysis Project 02/2022 to 07/2022

*Research Assistant, **NLP Group**, Massachusetts Institute of Technology, Advisor: Gary Becigneul*

- Developed an enhanced emotion extraction model that integrates image and text emotion retrieval, achieving a 9% improvement in prediction accuracy, particularly for the irony class, by classifying emotions into seven categories.
- Re-engineered the VistaNet model to enhance image information retrieval and implemented an image caption generation model with ResNet50 for attribute extraction, integrated these attributes with comments through BERT.
- Leveraged BERT for the fusion of dual text inputs to extract emotional content, incorporating a self-attention mechanism in the final layer for a precise amalgamation of image and text, optimizing emotional label prediction.

Face Spoofing Project 12/2021 to 05/2022

Research Assistant, Carnegie Mellon University, Advisor: Prof. Shlomo Ta'asan

- Implemented ViTGAN and SAGAN for facial spoof detection, enhancing facial restoration under varied lighting conditions, with a focus on analyzing generative and attention mechanisms for authentic representation.
- Crafted annotations to calibrate discriminator judgment, guiding generators in transitioning from dark to light scenarios, preserving facial integrity post-illumination adjustment.
- Demonstrated ViTGAN's superiority over SAGAN in capturing facial details, significantly improving face restoration from dark to well-lit conditions by leveraging its self-attention mechanism.

NEU Music Recommendation System Program 02/2023 to 07/2023

*Software Engineer, **NEUSoft,** Northeastern University. Advisor: Qiang Liu*

- Applied BERT and Retrieval algorithms using PyTorch and Spark for a sophisticated music recommendation system.
- Engineered platforms with Spring Boot and Vue, leveraging MySQL and MongoDB for backend data management.
- Enabled real-time, user-rating-based music recommendations by Kafka and Flume, with weighted blending for accuracy.

Skills

Programming Language: Python; Java; C; C++; HTML/CSS/JavaScript; GO

Machine Learning: PyTorch; TensorFlow; DeepSpeed; Transformers; LangChain; Scikit-learn; SpaCy; Pandas; NLTK

Systems & Framework: AWS EC2 & RDS, MySQL, Oracle, Linux; Slurm; MongoDB; Redis; LiteLLM