zxWenkai Li

wenkai.kyle.li@gmail.com | (412) 430-2334 | https://github.com/wenkai-li | https://wenkai-li.github.io

Research Interesting: Agentic system, LLM Social/Causal Reasoning, Responsible AI, Multimodal Machine Learning

Education

Carnegie Mellon University, Language Technology Institution, School of Computer Science 08/2023 to 12/2024

- Major 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

- Major GPA: 3.9/4.0
- 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

- **Wenkai Li,** Hyeonsu Kang, Roshni Kaushik, Xiaoyuan Wu, Maarten Sap, Koichi Onoue. "HypoVeil: A Hypothesis-Driven Pragmatic Inference-Time Control Framework for Privacy–Utility-Aware LLM-Agent Dialogue" under review in 2026 ICLR.
- Linus Tze En, Angela Ng, **Wenkai Li,** Lynnette Hui Xian Ng. "Singlish Then-And-Now: A Diachronic Study of Human- and LLM-Generated Texts" under review in 2026 AAAI.
- Xiaoyuan Wu, Roshni Kaushik, **Wenkai Li**, Koichi Onoue, "User Perceptions of Privacy and Helpfulness in LLM Responses to Privacy-Sensitive Scenarios" under review in 2026 IUI.
- **Wenkai Li***, Lynnette NG*, Andy Liu, Daniel Fried. "Measuring Fine-Grained Negotiation Tactics of Humans and LLMs in Diplomacy" under review in 2025 July ARR.
- 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" published 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 July ARR.
- **Wenkai Li*,** Jiarui Liu*, Andy Liu, Xuhui Zhou, Mona Diab, Maarten Sap. "BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data" published 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 & Work Experience

Social Aware Large Reasoning Model Self-evolve 10/2025 to Recent

Research Scientist Intern, *Fujitsu and Carnegie Mellon University. Advisor: Koichi, Niloofar Mireshghallah, Maarten Sap*

- Investigate social-cognitive pattern underlying Theory-of-Mind social reasoning in Large Reasoning Models, defining a taxonomy of meta-abilities and analyzing which reasoning trajectories consistently lead to successful outcomes.
- Develop a Process Reward Model (PRM) base self-evolution optimization loop combining Diversity Tree Search and Monte-Carlo Tree Search to explore diverse reasoning trajectories and iteratively refine the PRM that recognizes and rewards socially grounded reasoning.
- Conduct meta-ability validation across Theory-of-Mind datasets, comparing PRM-guided vs. baseline trajectories to

identify which reasoning compositions genuinely drive success while controlling for benchmark bias.

- Design targeted ability-shaping experiments using synthetic data with RL-based optimization methods, leveraging existing thinking trajectories and PRM traces to further guide and enhance the base model's social reasoning capabilities.

Multi-agent Scaling Law and Reward Base Refining 10/2025 to Recent

Research Scientist Intern, *Fujitsu and Carnegie Mellon University. Advisor: Koichi, Niloofar Mireshghallah, Maarten Sap*

- Study automatic role and task assignment in multi-agent LLM systems to enhance generalizability across domains, examining how well LLMs autonomously infer agent roles, decompose tasks, and dynamically adapt coordination strategies without manual prompt engineering.
- Investigate the scaling laws of automatically generated multi-agent collaboration, analyzing how agent count, task complexity, and coordination topology jointly affect performance–cost trade-offs.
- Design multi-granularity reinforcement learning strategies that optimize both macro-level outcomes and micro-level inter-agent cooperation, rewarding effective exchanges and penalizing redundant or conflicting behaviors to strengthen communication efficiency, credit assignment, and overall reasoning coherence.

Structural Translation: Multilingual Accessibility 08/2025 to Recent

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

- Formulated a new research problem on structural translation: evaluating how well current LLMs preserve HTML layout and accessibility elements in multilingual webpage translation tasks.
- Released a benchmark suite built from real-world LTI multilingual webpages, enabling systematic assessment of both textual fidelity and structural preservation across languages and models.
- Demonstrated structural degradation in leading LLMs under fine-grained instruction settings and introduced a self-refine translation agent that iteratively refines output using HTML parsing and visual preview tools.
- Conducted comprehensive human evaluations on both translation fidelity and structural integrity, establishing a well-defined and validated reference set to benchmark diverse models and prompting strategies.

Hypothesis-Driven Mental Model with LLM for solving Contextual Privacy 07/2025 to 09/2025

Research Scientist Intern*,* ***Data Security Lab****, Fujitsu, Advisor: Koichi Onoue, Maarten*

- Proposed HypoVeil, a hypothesis-driven inference-time control framework that integrates a mental-model belief store with Rational Speech Act (RSA)–based pragmatic reasoning, enabling LLM agents to reason about interlocutor beliefs, intentions, and contextual privacy norms during multi-turn conversation.
- Designed a dimension-aware mental-model that dynamically maintains natural-language hypotheses about the counterpart's knowledge, requests, and motives, guiding response selection under a quantified privacy–utility trade-off.
- Constructed VBench, a benchmark of contextual privacy scenarios, incorporating calibrated overlaps between sensitive and desired content, LLM-as-Judge scoring, and human-audited reliability checks for multi-turn evaluation.
- Demonstrated that HypoVeil consistently improves privacy preservation while retaining conversational helpfulness, showing clear Pareto-frontier gains and robustness across both closed- and open-source LLM backbones.

Large Language Model Negotiation with Diplomacy Setting Project 12/2024 to 07/2025

Research Assistant, ***Language Technology Institute****, Carnegie Mellon University. Advisors: Daniel Fried*

- 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.
- Developed an LLM-as-a-Judge evaluation pipeline to label fine-grained negotiation tactics in Diplomacy dialogues with reliability comparable to expert annotators.
- Conducted quantitative analyses linking rhetorical tactics to success, revealing strong correlations between socio-emotional and reasoning strategies and both short- and long-term game outcomes.
- Built predictive and alignment frameworks showing that supervised fine-tuning steers LLM negotiation behaviors toward human-like rhetorical distributions.
- Conduct comparative evaluation of human vs LLM negotiation styles, introducing an 8-dimensional ethos-pathos-logos

taxonomy and measuring model–human stylistic distances across multiple LLM backbones.

Context Engineer: Large Language Model Keep Secret Project 08/2024 to 12/2024

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

- Engineered a multi-agent framework (Extractor–Checker–Executor) with context engineering to modularize privacy reasoning, mitigating single-agent cognitive overload and improving contextual adherence.
- Performed systematic information-flow ablations to study how context visibility and inter-agent communication affect privacy–utility trade-offs, showing that checker model capacity critically determines privacy fidelity.
- Achieved 18–19% reduction in private leakage on contextual privacy (ConFaide and PrivacyLens) benchmarks with several LLMs backbones, maintaining public-content completeness.
- Conduct stage-wise analyses to quantify how engineered information flow enhances multi-agent transparency, error resilience, and contextual integrity preservation.

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

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

- Introduced BIG5-CHAT, a large-scale, human-grounded dialogue dataset capturing realistic Big Five personality expressions, bridging social interaction corpora with personality-annotated human text for authentic linguistic grounding.
- Developed a personality-steering framework (PSYCHSTEER) using the DExperts decoding mechanism and fine-tuned LLMs to embed personality traits through Supervised Fine-Tuning and Direct Preference Optimization, achieving stronger trait realism and intra-trait correlations than prompt-based methods.
- 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.
- Established the link between psycholinguistic personality traits and model cognition, providing quantitative evidence that human-grounded personality induction yields interpretable, psychologically consistent reasoning behaviors in LLMs.

Large Language Model Personality Steering Project 06/2024 to 12/2024

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

- Introduced GIST, the first large-scale multilingual AI terminology dataset with 5K terms from award-winning papers across 18 top-tier conferences (2000–2023), translated into five languages through a hybrid LLM + human framework validated by extensive crowdsourced evaluation.
- Developed a multi-stage pipeline combining term extraction, expert-verified translation, and LLM-assisted candidate selection, achieving higher accuracy and consistency than existing multilingual terminology resources.
- Conducted cross-lingual quality assessments using reference-based and human agreement metrics, and statistically confirmed significant improvements in translation fidelity and terminology correctness.
- Integrated GIST into machine translation workflows through prompting-based refinement and alignment methods, which improved domain terminology handling without the need for retraining.

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 02/2023 to 07/2023

***Software Engineer Intern, 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; Docker