Wenkai Li

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

Current Research Interests: AI safety, Security, Human-centered AI, Alignment & Post-training, Interpretability, Agentic

Education

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

- Major GPA: 3.9/4.0
- Advisor: Mona Diab, department head of LTI at CMU
- Relevant 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: 3.8/4.0
- Relevant Courses: Algorithm Analysis and Design; Discrete Mathematics; Data Mining Theories and Algorithms; Natural Language Processing; Web Development Programming Practice

UC Berkeley 08/2022 to 12/2022

- Visiting student, Major GPA: 3.9/4.0
- Relevant Courses: cs188: Introduce to Artificial Intelligence, cs 70: Discrete Mathematics and Probability Theory, Data 100: Principles and Techniques of Data Science

Publications

[10] 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.

[9] 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.

[8] 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.

[7] Wenkai Li*, Lynnette NG*, Andy Liu, Daniel Fried. "Measuring Fine-Grained Negotiation Tactics of Humans and LLMs in Diplomacy" under review in 2025 ARR.

[6] 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.

[5] Wenkai Li, Liwen Sun, Zhenxiang Guan, Xuhui Zhou, Maarten Sap. "1-2-3 Check: Enhancing Contextual Privacy in LLM via Multi-Agent Reasoning" accepted at IASEAI 2026.

[4] 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.

[3] Jiarui Liu, **Wenkai Li**, Zhijing Jin, Mona Diab. "Automatic Generation of Model and Data Cards: A Step Towards Responsible AI" published in 2024 NAACL.

[2] 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

[1] 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

Picked Research Experience

Large Reasoning Model Self-evolve 10/2025 to Recent

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

- Investigated reasoning strategies underlying Theory-of-Mind (ToM) social cognition in Large Reasoning Models by extracting special patterns and analyzing reasoning trajectories that consistently lead to successful social inferences.
- Developed a Process Reward Model (PRM) guided self-evolve loop combining Diversity Tree Search and Monte-Carlo Tree Search to iteratively refine the PRM, to recognize and reward structured, socially grounded reasoning processes.
- Explored token- and pattern-level mechanisms that elicit high-quality reasoning strategies—identifying linguistic and conceptual cues that correlate with coherent reasoning chains and improved ToM performance across models.
- Conducted comparative experiments and important patterns validation between PRM-guided and baseline reasoning trajectories, revealing how certain token-level patterns shape effective reasoning.

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

- Explored autonomous task and role definition in multi-agent LLM systems, investigating how a central planner can dynamically assign sub-agent workflows and infer optimal coordination structures without manually assigning.
- Analyzed performance–scaling dynamics of automatically generated multi-agent collaborations, examining how planner efficiency, task decomposition quality influence overall reasoning performance and coordination cost.
- Developed reinforcement learning–based optimization methods enabling the planner to refine collaboration strategies within sub-workflows, improving inter-agent communication, role alignment, and collective reasoning efficiency.

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.

Work Experience

Research Scientist Intern at Fujitsu Research of America **July 2025 - Recent**

Data&Security Lab, Mentor: Koichi Onoue Pittsburgh, PA

- Led HypoVeil, a hypothesis-driven inference-time control framework that integrates mental-model belief tracking and RSA-based pragmatic reasoning to enforce contextual privacy in multi-turn interactions.
- Led a multi-agent scaling study on automatic role assignment and coordination, developing RL-based planners to optimize collaboration structure, communication efficiency, and reasoning cost–performance trade-offs.
- Led a self-evolving reasoning project, designing PRM-guided search loops (DVTS + MCTS) to uncover Theory-of-Mind reasoning patterns and improve socially grounded reasoning trajectories in large reasoning models.
- Participate on other's work and prepared leading work on reward-model preference learning and reward hacking, and on analyzing meta-abilities emergence and neuron-level interpretability correlations in LLMs.

Software Engineer Intern at NEUSoft **May 2022 – Aug 2023**

Recommendation Team, Mentor: Qiang Liu Shenyang, Liaoning, China

- 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.

Professional Service

Conference Reviewer: ACL 2025, EMNLP 2025, NeurIPS 2025, ICLR 2026

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

Honors

- NEU School of Software Engineer scholarship – 2021
- NEU School of Software Engineer scholarship - 2022
- NEU Undergraduate University scholarship - 2022