

# MINGQIAN ZHENG

✉ mingqia2@andrew.cmu.edu    🌐 eeelisa.github.io    🎓 Google Scholar    ☎ 734-834-8955

## RESEARCH INTEREST

---

My research focuses on communication dynamics between humans and Large Language Models (LLMs), as well as interactions among multiple LLMs. I'm interested in how human-LLM communication patterns differ from human-human interactions, aiming to optimize these exchanges for improved human-AI collaboration. My work investigates LLM safety through refusal strategies, utility-safety trade-off evaluation, and multi-turn interactions, while also exploring multi-agent social simulations to understand LLMs' behavioral patterns and social reasoning capabilities in scenarios ranging from moral dilemma debates to non-cooperative interactions.

## EDUCATION

---

### Carnegie Mellon University

*Sep.2024 - Present*

*Ph.D. in Language and Information Technology*

Coursework: Advanced Natural Language Processing, Inference Algorithms for Language Modeling, Multimodal Machine Learning, Ethics, Safety, and Social Impact in NLP and LLMs

Advisors: Carolyn Rosé, Maarten Sap

### University of Michigan

*Sep.2022 - May.2024*

*Master in Survey and Data Science (Data Science Track),*

GPA: 4.0/4.0

Coursework: Natural Language Processing, Information Retrieval, Statistical Methods and Machine Learning, Machine Learning for Social Science

Advisor: David Jurgens, Yajuan Si

### New York University Shanghai

*Sep. 2018 - May 2022*

*Bachelor of Science in Mathematics*

GPA: 3.7/4.0

*Thesis: Modeling Short-term and Long-term Dynamics of User Intention for Missing-Not-At-Random Implicit Feedback in Recommendation Systems*

*Advisor: Shuyang Ling*

*Bachelor of Science in Data Science (Computer Science Track)*

*Thesis: User Intention Detection and Evaluation for Recommendation Systems*

*Advisors: Hongyi Wen, Oliver Marin*

Coursework: Machine Learning, Algorithms, Data Structures, Partial Differential Equations, Information Visualization, Databases, Numerical Analysis, Math Modeling, Probability and Statistics

## UNDER REVIEW

---

**Imperfectly Cooperative Human-AI Interactions: Comparing the Impacts of Human and AI Attributes in Simulated and User Studies**, Under Review

Myke C. Cohen\*, **Mingqian Zheng\***, Neel Bhandari, Hsien-Te Kao, Xuhui Zhou, Daniel Nguyen, Laura Cassani, Maarten Sap, Svitlana Volkova (\* denotes equal contributions)

## CONFERENCE PUBLICATIONS

---

**Let Them Down Easy! Contextual Effects of LLM Guardrails on User Perceptions and Preferences**, Findings of EMNLP 2025

**Mingqian Zheng**, Wenjia Hu, Patrick Zhao, Motahhare Eslami, Jena D. Hwang, Faeze Brahman, Carolyn Rose, Maarten Sap

*Media coverage: [Forbes]*

**Synthetic Socratic Debates: Examining Persona Effects on Moral Decision and Persuasion Dynamics**, EMNLP 2025

Jiarui Liu, Yueqi Song\*, Yunze Xiao\*, **Mingqian Zheng\***, Lindia Tjuatja, Jana Schaich Borg, Mona Diab, Maarten Sap (\* denotes equal contributions)

**When “A Helpful Assistant” Is Not Really Helpful: Personas in System Prompts Do Not Improve Performances of Large Language Models**, Findings of EMNLP 2024

Mingqian Zheng, Jiaxin Pei, Lajanugen Logeswaran, Moontae Lee, David Jurgens

Media coverage: [The Markup]

**Causally Modeling the Linguistic and Social Factors that Predict Email Response**, NAACL 2025

Yinuo Xu\*, Hong Chen\*, Sushrita Rakshit\*, Aparna Ananthasubramaniam\*, Omkar Yadav\*, **Mingqian Zheng\***, Michael Jiang\*, Lechen Zhang\*, Bowen Yi\*, Kenan Alkiek\*, Abraham Israeli\*, Bangzhao Shu\*, Hua Shen\*, Jiaxin Pei\*, Haotian Zhang\*, Miriam Schirmer\*, David Jurgens (\* denotes equal contributions)

**AWARDS**

---

**Rackham Graduate Student Research Grant at University of Michigan** Aug.2023  
**University Honors Scholar at New York University** May.2022  
**NYU Shanghai 2020 Recognition Award (Top 0.5%)** academic merit and community contributions Aug.2020

**RESEARCH EXPERIENCE**

---

**The Blablalab** May.2023 - Apr.2024

Advisor: Prof. David Jurgens at University of Michigan

- Built a pipeline for systematic evaluation of social roles in Large Language Models' system prompts
- Created a list of 162 roles covering 6 types of interpersonal relationships and 8 types of occupations, and showed that adding interpersonal roles in prompts consistently improves the models' performance over a range of questions through extensive analysis of 3 popular LLMs and 2457 questions
- Conceived and implemented multiple role-prediction methodologies, optimizing for the most effective role assignment in response to specific queries by evaluating frequency, similarity, and perplexity parameters

**Global Health and Population Project on Access to Care for Cardiometabolic Diseases** Nov.2022 - Apr.2024

Advisor: Prof. David Flood at University of Michigan

- Independently conducted data harmonization following the World Health Organization (WHO) STEPwise approach to noncommunicable disease risk factor surveillance (STEPS) using Stata
- Co-authored The Lancet Global Health publication on predicting diabetes prevalence and progress toward four WHO Diabetes Targets at the country level using a Bayesian statistical modeling approach

**User Intention Detection and Evaluation for Recommendation Systems** Feb.2022 - May.2022

Advisors: Prof. Hongyi Wen and Prof. Oliver Marin at NYU Shanghai

- Proposed an intention-aware recall metric along with a statistical mechanics of user intention extraction based on the MovieLens 1M dataset
- Verified the reliability of the metric in evaluating recommendation models by comparing the scores and performances of four DNN and RNN models

**INTERNSHIP EXPERIENCE**

---

**Natural Language Processing Research Intern(Core R&D Group) at iFlytek** Jul.2021 - Oct.2021

- Applied Recurrent Neural Networks (RNN), sequence labelling and supervised classification methods to implement word-level language identification in a code-switching dataset containing English and Spanish
- Trained the model to achieve accuracy 86% based on the cleaned dataset of 3332 English-Spanish multilingual tweets and the model was used as the baseline model for the future product

**COMMUNITY SERVICES**

---

**Conference Reviewer** ACL Rolling Review 2024-2025, NeurIPS 2025  
**Book Reviewer** Synthesis Lectures on Human Language Technologies