Lechen Zhang

Ann Arbor, MI | +1 734 834 8529 | leczhang@umich.edu
Google Scholar | GitHub | Twitter | Homepage

Education

University of Illinois Urbana-Champaign *Aug. 2025 – May. 2029 (Expected)*
PhD in Computer Science

- **Incoming 25Fall Student**

University of Michigan, Ann Arbor *Aug. 2022 – May. 2024*
Master in Information Science | ***GPA: 4.00/4.00*** | ***Distinguished Thesis***

- **Advisor**: Prof. David Jurgens and Prof. Lu Wang
- **Related coursework:** Applied Data Science (A+), Information Retrieval (A+), Game Theory (A+), Math for Data Science (A+), Big Data Analytics (A+), NLP Algorithm (A), Machine Learning (A)

Shanghai Jiao Tong University *Sep. 2019 – Aug. 2023*
Bachelor in Electrical and Computer Engineering

- **Related coursework:** Computer Vision (A), Computer Architecture (A), Intro to Linguistics (A)

Publications

** indicates equal contribution*

Peer-Reviewed Papers

[8] **FactBench: A Dynamic Benchmark for In-the-Wild Language Model Factuality Evaluation**
Farima Fatahi Bayat, **Lechen Zhang**, Sheza Munir, Lu Wang
ACL 2025 | Main Contributor [arXiv] [Code] [Twitter]

[7] **Towards Global AI Inclusivity: A Large-Scale Multilingual Terminology Dataset (GIST)**
Jiarui Liu*, Iman Ouzzani*, Wenkai Li*, **Lechen Zhang**, Tianyue Ou, Houda Bouamor, Zhijing Jin, Mona Diab
ACL 2025 Findings [arXiv]

[6] **Causally Modeling the Linguistic and Social Factors that Predict Email Response**
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 (*Randomized Author Order*)
NAACL 2025 | Main Contributor [ACL Anthology]

[5] **You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments**
Bangzhao Shu*, **Lechen Zhang***, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, David Jurgens
NAACL 2024 Oral | Project Leader [ACL Anthology] [arXiv] [Slides] [Code] [Twitter]

Papers Under Review & Preprints

[4] **SPRIG: Improving Large Language Model Performance by System Prompt Optimization**
Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, David Jurgens
ARR 2025 Under Review | Project Leader [arXiv] [Slides] [Code] [Twitter]

[3] **Latent Geographies: Joint Embeddings of Text and Visual Cues for Social Media Geolocation**
Lechen Zhang*, Abraham Israeli*, Rohan Raju, David Jurgens
ARR 2025 Under Review | Project Leader

[2] **Real or Robotic? Assessing Whether LLMs Accurately Simulate Qualities of Human Responses in Dialogue**
Jonathan Ivey*, Shivani Kumar*, Jiayu Liu*, Hua Shen*, Sushrita Rakshit*, Rohan Raju*, Haotian Zhang*, Aparna Ananthasubramaniam*, Junghwan Kim*, Bowen Yi*, Dustin Wright*, Abraham Israeli*, Anders Giovanni Møller*, **Lechen Zhang***, David Jurgens (*Randomized Author Order*)
ARR 2025 Under Review | Project Leader [arXiv] [Code] [Twitter]

[1] **VeriFact: Enhancing Long-Form Factuality Evaluation with Refined Fact Extraction and Reference Facts**
Xin Liu, **Lechen Zhang**, Sheza Munir, Yiyang Gu, Lu Wang
ARR 2025 Under Review | Main Contributor [arXiv]

Research Experience

Concluded Projects

Improving LLMs' general performance by System Prompt Optimization *Feb. 2024 – Present*
Advisor: David Jurgens *University of Michigan*

- Design an edit-based genetic system prompt optimizer *SPRIG* that generally improves LLM performance across 47 diverse benchmarks.
- Discover strong generalization capability of system prompt optimization across tasks, models, languages, and even out-of-domain challenges, and its complementary effect with existing task-specific optimizers.
- Develop new RL strategies to efficiently explore and expand the design space of system prompts.
- Lead the whole project independently, completing all aspects from research ideation to paper writing.

Factuality Evaluation pipeline and benchmark in real-world scenarios *May. 2024 – Nov. 2024*
Advisor: Lu Wang *University of Michigan*

- Develop a retrieval-based factuality evaluation pipeline for long-form text that is more fine-grained, efficient and aligns better with human than existing work.
- Build a benchmark of prompts that are factually challenging to LLMs by filtering LMSYS-1M dataset, clustering representative prompts, and selecting based on their scores on the designed evaluation pipeline.
- Lead experiments on open-source models, reproduce 3 existing studies as baselines, and implement parallel optimization for a 10x speedup.
- Contribute extensively to paper writing and post-submission tasks, including drafting key sections, analyzing results, creating visuals, managing code repository, and preparing rebuttals.

Assessment of LLM Simulation Ability of Human Responses in Dialogue *Jul. 2024 – Oct. 2024*
Advisor: David Jurgens *University of Michigan*

- Implement 15 evaluation metrics for LLM simulation quality across lexical, syntactic, semantic, and style features.
- Lead the collection of 50 instruction prompts and generate 1M dialogue simulation results from 9 LLMs.
- Set up annotation platform for the whole lab to collect human annotations as a baseline.
- Lead the project as the main contributor to coding, paper writing and post-submission tasks.

Modeling Intent, Expectation, and Responsiveness in Email Conversations *Mar. 2024 – Jun. 2024*
Advisor: David Jurgens *University of Michigan*

- Preprocess raw Email data and build an email relationship network to sample and construct a dataset for analysis.
- Evaluate LLM's ability to infer Email Intent by fine-tuning RoBERTa and running zero-shot inference on Llama-3.
- Serve as the main contributor to annotating, coding, paper writing and post-submission tasks.

Robustness of LLMs' personality under Psychometric Instruments *Sep. 2023 – Dec. 2023*
Advisor: David Jurgens *University of Michigan*

- Build evaluation dataset and metrics that measures the robustness of various LLMs' personalities under spurious prompt variation and rephrased statements, and evaluate on 17 different LLMs.
- Experiment the personality and robustness shifts under different conditions, such as injecting personalities through prompts, and fine-tuning LLMs (Llama2, Flan-T5, etc.) on various corpora (Bible, 4chan, r/Donald, etc.).
- Lead the project and contribute to most coding, experiments, writing, and post-submission tasks such as the rebuttal, code repository, Twitter thread, and related presentations.

Ongoing Projects

Optimized Data Selection and Mixture for Scalable and Efficient LLM Post-Training *Oct. 2024 – Present*
Advisor: Lu Wang and Wei Hu *University of Michigan*

- Explore data heterogeneity and develop methods to automatically discover "skills" for better data mixture laws.
- Experiment with model merging methods as a representation for efficient SFT data selection and compare with existing data mixture approaches.
- Evaluate OLMo checkpoints to investigate the emergence of capabilities in LLMs and identify critical transition points.

Enhancing LLMs' Multimodal Capabilities in Identifying National Culture Difference *Oct. 2024 – Present*
Advisor: David Jurgens *University of Michigan*

- Develop contrastive learning methods that train multimodal LLMs to identify the national culture difference in Twitter and Mastodon data.
- Mentor an undergraduate student and serve as the main contributor to coding and paper writing.

Presentations

NAACL 2024 Oral (Mexico City) – *You don't need a personality test to know these models are unreliable: Assessing the Reliability of Large Language Models on Psychometric Instruments.* [Slides] [ACL Anthology]

Professional Services

Conference Reviewer

- NAACL 2025
- EMNLP 2024 (***Outstanding Reviewer***)

Skills

Programming: Python, C/C++, C#, Java, Go, SQL, MATLAB, R, Kotlin, LaTeX

Frameworks: PyTorch, Tensorflow, Transformers, Accelerate, DeepSpeed, PEFT, NLTK, Scikit-Learn

Languages: Chinese (Native), English (Fluent), Japanese (Basic)