YULING GU

yuling.gu@nyu.edu | yulinggu-cs.github.io | @gu_yuling | yuling-gu | Google Scholar

EDUCATION

Ph.D. in Data Science

New York University, New York, NY, USA

September 2025 - Present

- Ph.D. student at Center for Data Science
- Center for Data Science Fellowship (5 years)
- Coursework taken/ in progress: Introduction to Data Science for PhD Students, Probability and Statistics for Data Science, Big Data

Master of Science

University of Washington, Seattle, WA, USA

September 2020 – March 2022

- Master's in Electrical and Computer Engineering (Cumulative GPA: 4.00/4.00)
- Relevant Coursework: Statistical Learning, Continuous-Space Language Processing, Syntax For Computational Linguistics, Computational Cognitive And Neural Modeling, Neural Computation And Engineering, Probability and Random Processes, Mathematical Foundations of Systems Theory

Bachelor of Arts, *summa cum laude*

New York University, New York, NY, USA

August 2016 - May 2020

- Bachelor of Arts, *summa cum laude* (Cumulative GPA: 3.991/4.00)
- Double major: Computer Science with high honors, Language and Mind (Joint major in linguistics, psychology and philosophy); Minor: Mathematics
- Honors Thesis: "Towards detecting temporal relations implicitly conveyed in text"
  Advisor: Prof. Ernest Davis
- Relevant Coursework: Machine Learning for Language Understanding, Introduction to Machine Learning, Artificial Intelligence, Natural Language Processing, Theory of Computation, Numerical Computing, Operating Systems, Basic Algorithms, Computer Systems Organization, Data Structures, Discrete Mathematics, Linear Algebra, Calculus 3, Probability & Statistics, Statistics for The Behavioral Sciences, Sound and Language, Grammatical Analysis, Introduction to Semantics, Neural Bases of Language, Minds and Machines, Logic, Language and Mind, Perception, Cognition

RESEARCH EXPERIENCE

Predoctoral Young Investigator

Allen Institute for AI

April 2022 – August 2025 | Seattle, WA

- Contributed to collaborative efforts building open source models such as OLMo (Best Theme Paper at ACL 2024), OLMo 2, OLMo 3, OLMoE, and TÜLU 3, by leading efforts to build and iterate on the evaluation framework/suite.
- Created SimpleToM, a novel Theory-of-Mind (ToM) dataset that exposes the gap between explicit ToM inference and implicit ToM application in models.
- Proposed OLMES, a standard for reproducible LLM evaluations that is open, practical, and completely documented.
- Built Digital Socrates, a system that provides high-quality, nuanced automatic evaluation of explanations from models for the first time, filling an important gap in evaluation tools for the community.
- Proposed a way of materializing LLMs' mental models of everyday things (like an egg, tree, or bicycle) for the first time, revealed the incoherence in mental models of SOTA LLMs, and developed a neuro-symbolic method to improve them.
- Built the DREAM model, the first work to show that adding focused elaborations about a situation using social constructs (e.g., motivations) can improve LLMs' reasoning about the situations.

- Demonstrated the effectiveness of using "mental models" of situations for figurative language understanding (Led a team and built a winning system - joint first place for Figurative Language Understanding Shared Task at EMNLP 2022).

Research intern

Allen Institute for AI

Summer 2021 – Fall 2021 | Seattle, WA

- Supervised by Dr. Bhavana Dalvi and Dr. Peter Clark. Proposed an approach for building "mental models" of situations given in input text. Demonstrated how this technique enables language models to better understand and answer situation-based questions (improved task performance on various tasks such as ETHICS, CODAH, and Social IQA).

Undergraduate research assistant

New York University

Summer 2018 – Spring 2020 | New York, NY

- **Honors Thesis Project: Detecting temporal relations in text**
  (Spring 2019 - Spring 2020)
  Supervised by Prof. Ernest Davis. Used various classifiers, word and sentence representations, as well as linguistics theories to automatically detect temporal relations implicitly conveyed in texts (different levels: from single event description to multiple sentences); Culminated in honors thesis "Towards detecting temporal relations implicitly conveyed in text."
- **Can dependency parsing help event extraction in text?**
  (Fall 2019 - Spring 2020)
  Supervised by Prof. Ralph Grishman. Investigated the contribution of information from dependency parsing, Named Entity (NE) tagging, and Part Of Speech (POS) tagging in event extraction, beyond a baseline that uses pretrained BERT sentence representation.
- **Integrated Customization Environment for Information Extraction (ICE)**
  (Summer 2019)
  Supervised by Prof. Ralph Grishman. Experimented with different classifiers, together with grammatical linguistics insights, to automatically distinguish prepositional phrases as adjuncts or arguments (achieved 88% accurate prediction of the adjunct/argument distinction using linguistics theories alone).
- **Termolator: A terminology extraction system (Open source tool)**
  (Summer 2018 - Fall 2018)
  Supervised by Prof. Adam Meyers. Refined the English Termolator's distributional metrics; Further developed the Chinese Termolator; Integrated past 5 years' developments to unify the two systems (GitHub: my contributions integrated to full system on July 2020).
- **Independent study project: Commonsense Reasoning**
  (Summer 2018)
  Supervised by Prof. Ernest Davis. Looked into English-Chinese Machine Translation failures; Designed Winograd schemas and compile pronoun disambiguation problems; Worked on Toward Annotating Commonsense Inferences in Text (TACIT) annotation.

Research intern

Institute for Infocomm Research, A*STAR

Winter 2014 – Spring 2021 | Singapore, Singapore

- Characterized Singaporean, American, and British English acoustic and pronunciation patterns in children's speech using unsupervised clustering (supervised by Dr. Nancy F. Chen).
- Compared Chinese tone perception in Singaporean and native Chinese Mandarin speakers; Investigated tone in whispered Mandarin (jointly supervised by Dr. Boon Pang Lim and Dr. Nancy F. Chen).

OTHER WORK EXPERIENCE

Grader

New York University

Spring 2019, Fall 2019 | New York, NY

- Grader for Artificial Intelligence course at Courant Institute of Mathematical Sciences (CIMS) under Prof. Ernest Davis. (Fall 2019)
- Grader for Basic Algorithms course at Courant Institute of Mathematical Sciences (CIMS) under Prof. Victor Shoup. (Spring 2019)

PUBLICATIONS

[1] **Yuling Gu**, Oyvind Tafjord, Hyunwoo Kim, Jared Moore, Ronan Le Bras, Peter Clark, and Yejin Choi (2026). "SimpleToM: Exposing the Gap between Explicit ToM Inference and Implicit ToM Application in LLMs". ICLR 2026.

[2] Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, **Yuling Gu**, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A Smith, and Hannaneh Hajishirzi (2025). "OLMo 3". arXiv 2025.

[3] David Heineman, Valentin Hofmann, Ian Magnusson, **Yuling Gu**, Noah A Smith, Hannaneh Hajishirzi, Kyle Lo, and Jesse Dodge (2025). "Signal and Noise: A Framework for Reducing Uncertainty in Language Model Evaluation". NeurIPS 2025.

[4] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, **Yuling Gu**, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi (2025). "TÜLU 3: Pushing Frontiers in Open Language Model Post-Training". COLM 2025.

[5] **Yuling Gu**, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi (2025). "OLMES: A Standard for Language Model Evaluations". Findings of NAACL 2025.

[6] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, **Yuling Gu**, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A Smith, and Hannaneh Hajishirzi (2025). "2 OLMo 2 Furious". COLM 2025.

[7] Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, **Yuling Gu**, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi (2025). "OLMoE: Open Mixture-of-Experts Language Models". ICLR 2025.

[8] Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and **Yuling Gu** (2024). "WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models". LREC-COLING 2024.

[9] Tianyi Zhang, Li Zhang, Zhaoyi Hou, Ziyu Wang, **Yuling Gu**, Peter Clark, Chris Callison-Burch, and Niket Tandon (2024). "PROC2PDDL: Open-Domain Planning Representations from Texts". ACL 2024 • The 2nd Workshop on Natural Language Reasoning and Structured Explanations.

[10] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, **Yuling Gu**, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi (2024). "OLMo: Accelerating the Science of Language Models". ACL 2024. (Best Theme Paper Award)

[11] **Yuling Gu**, Oyvind Tafjord, and Peter Clark (2024). "Digital Socrates: Evaluating LLMs through explanation critiques". ACL 2024.

[12] Kavel Rao, Liwei Jiang, Valentina Pyatkin, **Yuling Gu**, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi (2023). "What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations". Findings of EMNLP 2023.

[13] Danica Dillion, Niket Tandon, **Yuling Gu**, and Kurt Gray (2023). "Can AI language models replace human participants?". Trends in Cognitive Sciences 2023.

[14] **Yuling Gu**, Bhavana Dalvi, and Peter Clark (2023). "Do language models have coherent mental models of everyday things?". ACL 2023.

[15] **Yuling Gu** (2022). "Measure More, Question More: Experimental Studies on Transformer-based Language Models and Complement Coercion". arXiv 2022.

[16] Bingchen Zhao*, **Yuling Gu***, Jessica Zosa Forde, and Naomi Saphra (2022). "One Venue, Two Conferences: The Separation of Chinese and American Citation Networks". NeurIPS 2022 • AI Cultures Workshop.

[17] **Yuling Gu**, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi, and Peter Clark (2022). "Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE". EMNLP 2022 • The Third Workshop on Figurative Language Processing. (Joint first place for the Shared Task)

[18] **Yuling Gu**, Bhavana Dalvi, and Peter Clark (2022). "DREAM: Improving Situational QA by First Elaborating the Situation". NAACL 2022.

[19] **Yuling Gu** and Nancy F. Chen (2022). "Large-Scale Acoustic Characterization of Singaporean Children's English Pronunciation". arXiv 2022.

[20] **Yuling Gu** (2021). "Transformer-based language models and complement coercion: Experimental studies". ACL-IJCNLP 2021 • UnImplicit: The First Workshop on Understanding Implicit and Underspecified Language.

[21] **Yuling Gu** (2020). "Towards detecting temporal relations implicitly conveyed in text". Undergraduate honors thesis.

[22] **Yuling Gu** and Nancy F. Chen (2020). "Characterization of Singaporean Children's English: Comparisons to American and British Counterparts using Archetypal Analysis". Interspeech 2020.

[23] **Yuling Gu** and Nancy F. Chen (2019). "Large-scale acoustic characterization of mid-low vowels across American, British, and Singaporean children". *The Journal of Acoustical Society of America*, *Volume 146*, *Issue 4*. 178th Meeting of the Acoustical Society of America.

[24] **Yuling Gu** and Nancy F. Chen (2019). "Acoustic characterization of Singaporean children's English with American and British counterparts: A case study on approximants". *The Journal of Acoustical Society of America*, *Volume 146*, *Issue 4*. 178th Meeting of the Acoustical Society of America.

[25] **Yuling Gu** and Nancy F. Chen (2019). "Acoustic Characterization of Singaporean Children's English: Comparisons to American and British Counterparts". ACL 2019 • Widening NLP workshop.

[26] **Yuling Gu**, Boon Pang Lim, and Nancy F. Chen (2016). "Perception of tone in whispered Mandarin sentences: the case for Singapore Mandarin". Interspeech 2016.

AWARDS/HONORS

University of Washington
- The Graduate School Top Scholar Research Assistantship Award (first year)

New York University
- Data Science Fellowship Award (Fall 2025 – Spring 2030)
- Phi Beta Kappa Honor Society (2020)
- Faculty Memorial Award (Dean's Award Spring 2020)
- Computer Science Prize for Academic Excellence in the Honors Program (Dean's Award Spring 2020)
- University Honors Scholar (Spring 2020)
- Four time recipient of Dean's Undergraduate Research Fund (Spring 2018 Individual Research Grant, Spring 2019 Individual Research Grant & 2 Conference Grants)
- Bernice Nachman Marlowe Scholarship (Fall 2018 – Spring 2020)
- Steffi Berne Scholarship (Fall 2018 – Spring 2020)
- Dean's List for Academic Year (all 4 academic years)
- Presidential Honors Scholars (Fall 2017 – Spring 2020)
- Women In Science (WINS) Scholarship (Fall 2017 – Spring 2020)

External Grants
- Acoustical Society of America travel grant (for 178th Meeting)
- Widening NLP workshop at ACL travel grant (for ACL 2019)
- ISCA travel grant (for Interspeech 2016)

SERVICE

Reviewer for COLM 2026, ACL ARR 2024 October, ACL ARR 2023 December, MP2 Workshop at NeurIPS 2023, AIJ 2023, AAAI 2022, and Workshop on Figurative Language Processing at EMNLP 2022.

OTHER ACTIVITIES

Allen Institute for AI

- AI2 Hackathon (August 11-13, 2021; August 10-12, 2022)

New York University

- NYU CAS alumni-student debate. The Motion: "The Benefits of the Development of Artificial Intelligence Outweigh the Harms." (October 26, 2019)
- Represented NYU at 2019 Grace Hopper Celebration of Women in Computing in Orlando, FL, USA
- Women in Science visibility committee: liaison (Fall 2017 – Spring 2020)
- Presidential Honors Scholars Program (Fall 2017 – Spring 2020)
- League of Linguistics interest group (Fall 2017 – Spring 2020)
- Dean's Service Honors Corps (Fall 2017 – Spring 2020)
- Minority And Philosophy chapter (student-organized interest group) (Spring 2017 – Spring 2020)