YANJUN GAO

Assistant Professor
Department of Biomedical Informatics
University of Colorado, Anschutz Medical Campus

https://serenayj.github.io/
yanjun.gao@cuanschutz.edu

EDUCATION

Ph.D. Computer Science and Engineering
Pennsylvania State University, State College, PA, USA   2014 – 2021
Advisor: Dr. Rebecca J. Passonneau
Field of Study: Natural Language Processing (NLP)
THESIS - *Analysis of Text to Identify, Represent and Group Distinct Propositions*

Bachelor of Management, Electronic Commerce
Harbin Finance University, Harbin, China   2010 – 2014
GPA Ranking: Top 1%
China National Scholarship

Minor, English
Heilongjiang University, Harbin, China   2010 – 2014
Second foreign language: Japanese

PROFESSIONAL EXPERIENCE

Postdoctoral Research Associate   August 2021 – August 2024
University of Wisconsin Madison, Madison, WI
- Authored a systematic review on clinical natural language processing (NLP) tasks, developed a novel annotation framework for electronic health records and designed a new suite of tasks for clinical NLP.
- Pioneered the first clinical NLP benchmark for diagnostic reasoning, establishing baseline results using the leading NLP methods including fine-tuning large language models (LLMs), and innovated a novel knowledge graph prompting method for LLMs in diagnosis prediction, mentored PhD students to design human and automated evaluation metrics for diagnosis prediction.
- Designed an automated prompt engineering framework for healthcare applications and published papers detailing these advancements.
- Served as NLP consultant to the University of Wisconsin Hospital (UW Health) operational team to design human-factor prompting and evaluation framework for LLM healthcare use cases.

Research Assistant, Pennsylvania State University, State College, PA

NLP Lab, Department of Computer Science,   2017 – 2021
- Developed "PyrEval", an innovative method and software package for summarization evaluation, achieving higher human correlation than conventional metrics such as ROUGE.
- Developed graph-based methods for discourse connective predictions and decomposition of complex sentences into individual propositions
- Collaborated with researchers in education and collected data from real-world classrooms on student writings, conducted reliability studies on automated pyramid summarization evaluation in student writing and authored papers to publish these findings.

School of Information Science and Technology   2015 – 2016
- Designed and implemented a multi-user system for collaborative information searching, sharing, and decision-making in an online environment. Conducted pilot studies, performed quantitative and qualitative analysis and published results.

Research Intern, Kwai Inc (Kuaishou Technology)   May 2019 – August 2019
Silicon Valley AI Lab, Palo Alto, CA

- Designed and implemented an advanced multi-modal pipeline to fuse video-understanding neural networks and text encoders, facilitating video-language understanding and interaction in temporal grounding tasks. Conducted comprehensive experiments and successfully published the results.

GRANTS

National Library of Medicine K99/R00 Pathway to Independence Award   2023 – 2027
National Institutes of Health, "Developing and Evaluating Multi-Modal Clinical Diagnostic Reasoning Models for Automated Diagnosis Generation", Amount: $ 874,800

JOURNAL PUBLICATIONS

- Afshar M, **Gao Y**, Gupta D, Croxford E, Demner-Fushman D. On the role of the UMLS in supporting diagnosis generation proposed by Large Language Models. *J Biomed Inform.* 2024 Aug 13;157:104707. doi: 10.1016/j.jbi.2024.104707. Epub ahead of print. PMID: 39142598.

- Afshar, Majid, **Yanjun Gao**, Graham Wills, Jason Wang, Matthew M. Churpek, Christa J. Westenberger, David T. Kunstman, Joel E. Gordon, Frank J. Liao, and Brian W. Patterson. "Prompt Engineering GPT-4 to Answer Patient Inquiries: A Real-Time Implementation in the Electronic Health Record across Provider Clinics." medRxiv (2024): 2024-01.

- **Gao, Yanjun**, Diwakar Mahajan, Ozlem Uzuner, and Meliha Yetisgen. "Clinical natural language processing for secondary uses." Journal of Biomedical Informatics (2024): 104596.

- **Gao, Yanjun**, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W. Patterson, Matthew M. Churpek, Timothy Miller, Dmitriy Dligach, Majid Afshar. "Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction". *Journal of American Medical Informatics Association.* 2023. (Under Review).

- **Gao, Yanjun**, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, Ozlem Uzuner, and Majid Afshar. "Progress Note Understanding—Assessment and Plan Reasoning: Overview of the 2022 N2C2 Track 3 shared task." *Journal of Biomedical Informatics* (JBI 2023): 104346.

- **Gao, Yanjun**, Dmitriy Dligach, Timothy Miller, John Caskey, Brihat Sharma, Matthew M. Churpek, and Majid Afshar. "DR. BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing." *Journal of Biomedical Informatics* 138 (JBI 2023): 104286.

- **Gao, Yanjun**, Dmitriy Dligach, Leslie Christensen, Samuel Tesch, Ryan Laffin, Dongfang Xu, Timothy Miller, Ozlem Uzuner, Matthew M. Churpek, and Majid Afshar. "A Scoping Review of Publicly Available Language Tasks in Clinical Natural Language Processing." *Journal of the American Medical Informatics Association (JAMIA, 2022).*

- Yetisgen, Meliha, Ozlem Uzuner, **Yanjun Gao**, and Diwakar Mahajan. "Call for papers: Special issue on clinical natural language processing for secondary use applications." *Journal of Biomedical Informatics 133 (JBI, 2022)*: 104152-104152.

- Davies, Patricia Marybelle, Rebecca Jane Passonneau, Smaranda Muresan, and **Yanjun Gao**. "Analytical techniques for developing argumentative writing in STEM: A pilot study." *IEEE Transactions on Education 65, no. 3 (2021)*: 373-383.

CONFERENCE AND OTHER PUBLICATIONS

- **Gao, Yanjun**, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A. Miller, Danielle Bitterman, Matthew Churpek, and Majid Afshar. "When Raw Data Prevails: Are Large Language Model Embeddings Effective in Numerical Data Representation for Medical Machine Learning Applications?." Findings of Empirical Methods in Natural Language Processing (*EMNLP* 2024).

- Croxford, Emma, **Yanjun Gao**, Brian Patterson, Daniel To, Samuel Tesch, Dmitriy Dligach, Anoop Mayampurath, Matthew M. Churpek, and Majid Afshar. "Development of a Human Evaluation Framework and Correlation with Automated Metrics for Natural Language Generation of Medical Diagnoses." American Medical Informatics Association (*AMIA* 2024).

- Chen, Xin, Hanxian Huang, **Yanjun Gao**, Yi Wang, Jishen Zhao, and Ke Ding. "Learning to Maximize Mutual Information for Chain-of-Thought Distillation." Findings of Association of Computational Linguistics (*ACL* 2024).

- **Gao, Yanjun**, Ruizhe Li, Emma Croxford, Samuel Tesch, Daniel To, John Caskey, Brian W Patterson, Matthew M Churpek, Timothy Miller, Dmitriy Dligach, Majid Afshar. "Large Language Models and Medical Knowledge Grounding for Diagnosis Prediction". medRxiv 2023.11.24.23298641; doi: https://doi.org/10.1101/2023.11.24.23298641

- **Gao, Yanjun**, Ruizhe Li, John Caskey, Dmitriy Dligach, Timothy Miller, Matthew M. Churpek, and Majid Afshar. "Leveraging A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction." arXiv preprint arXiv:2308.14321 (2023).

- **Gao, Yanjun**, Dmitriy Dligach, Timothy Miller, Matthew MM Churpek, and Majid Afshar. "Overview of the Problem List Summarization (ProbSum) 2023 Shared Task on Summarizing Patients' Active Diagnoses and Problems from Electronic Health Record Progress Notes." *In The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 461–467, Toronto, Canada. Association for Computational Linguistics. (BioNLP, 2023)*

- Brihat Sharma, **Yanjun Gao**, Timothy Miller, Matthew M. Churpek, Majid Afshar and Dmitriy Dligach. "Multi-Task Training with In-Domain Language Models for Diagnostic Reasoning." *In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 78–85, Toronto, Canada. Association for Computational Linguistics. (ClinicalNLP, 2023)*

- Zhou, Weipeng, Meliha Yetisgen, Majid Afshar, **Yanjun Gao**, Guergana Savova, and Timothy Miller. "Improving Model Transferability for Clinical Note Section Classification Models Using Continued Pretraining." *In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 125–130, Toronto, Canada. Association for Computational Linguistics. (ClinicalNLP, 2023)*

- **Gao, Yanjun**, Caskey, John, Miller, Timothy, Sharma, Brihat, Churpek, Matthew, Dligach, Dmitriy, and Majid Afshar. "Tasks 1 and 3 from Progress Note Understanding Suite of Tasks: SOAP Note Tagging and Problem List Summarization" (version 1.0.0). PhysioNet (2022). https://doi.org/10.13026/wks0-w041.

- **Gao, Yanjun**, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew MM Churpek, and Majid Afshar. "Summarizing Patients' Problems from Hospital Progress Notes Using Pre-trained Sequence-to-Sequence Models." *In Proceedings of the 29th International Conference on Computational Linguistics, pp. 2979-2991. (COLING, 2022).*

- **Gao, Yanjun**, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. "Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding." *Proceedings of the 13th Language Resources and Evaluation Conference. (LREC, 2022).*

- **Gao, Yanjun**, (2021). Analysis of Text to Identify, Represent and Group Distinct Propositions. *Thesis.*

- **Gao, Yanjun**, Kenneth Ting-hao Huang, Rebecca J. Passonneau. "ABCD: A Graph Framework to Convert Complex Sentences to a Covering Set of Simple Sentences". *59th Annual Meeting of the Association for Computational Linguistics (ACL, 2021).*

- **Gao, Yanjun**, Lulu Liu, Jason Wang, Rui Zhang. "EVOQUER: Enhancing Temporal Grounding with Video-Pivoted Back Query Generation". *4th Workshop on Visually Grounded Interaction and Language (ViGIL, 2021)*.

- **Gao, Yanjun**, Rebecca J. Passonneau. "Automated Assessment of Quality and Coverage of Ideas in Students' Source-based Writing".*22nd International Conference on Artificial Intelligence in Education (AIED, 2021)*

- **Gao, Yanjun**, Kenneth Ting-hao Huang, Rebecca J. Passonneau. "Learning Clause Representation from Dependency-Anchor Graphs for Discourse Connective Prediction". *15th Workshop on Graph-Based Natural Language Processing (TextGraphs, 2021)*.

- **Gao, Yanjun**, Xin Chen, Huayan Wang. "Video Pivoted Machine Translation Assisted Temporal Grounding." *Patent Filed in China, 2021.*

- **Gao, Yanjun**, Chen Sun, Rebecca J. Passonneau. "Automatic Pyramid Summarization Evaluation". *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL). 2019.*

- **Gao, Yanjun**, Alex Driban, Brennan Xavier McManus, Elena Musi, Patricia Davies, Smaranda Muresan, and Rebecca J. Passonneau. "Rubric Reliability and Annotation of Content and Argument in Source-Based Argument Essays." *In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2019), pp. 507-518.* 2019.

- **Gao, Yanjun**, Patricia M. Davies, and Rebecca J. Passonneau. "Automated content analysis: A case study of computer science student summaries." *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications (BEA, 2018).*

- **Gao, Yanjun**, Andrew Warner and Rebecca J. Passonneau. "PyrEval: An Automated Method for Summary Content Analysis", *11th edition of the Language Resources and Evaluation Conference (LREC, 2018)*

- **Gao, Yanjun**, Madhu Reddy and Bernard J.Jansen. "ShopWithMe!: Collaborative Information Searching and Shopping for Online Retail". *Proceedings of the 50th Hawaii International Conference on System Sciences (HICSS). 2017.*

- **Gao, Yanjun**, Madhu Reddy, and Bernard J. Jansen. "Shop together, search together: Collaborative e-commerce." *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems. 2016.*

DISTINCTIONS

- Biomedical Informatics Year-in-Review (nomination), American Medical Informatics Association (AMIA) Annual Symposium 2023.

- Outstanding Reviewer, npj Digital Medicine, Nature 2023.

- Trainee Outstanding Research Awards at Department of Medicine, University of Wisconsin Madison 2023.

- Best Clinical Research at Department of Medicine Annual Research Day, University of Wisconsin Madison 2022.

- Oral Presentation, Association of Computational Linguistics (ACL) 2021.

- Oral Presentation, Conference on Computational Natural Language Learning (CoNLL) 2019.

- Best Poster in New York Academy of Sciences NLP, Dialog and Speech Symposium (NYAS NDS) 2019

- Oral Presentation, Hawaii International Conference on System Sciences (HICSS) 2017.

- National Scholarship, Ministry of Education in China 2013 – 2014.

GUEST EDITOR

Journal of Biomedical Informatics

ORGANIZING COMMITTEE

International Workshop on Graph-based Natural Language Processing (TextGraphs 2022, 2024)
Biomedical Natural Language Processing Workshop (BioNLP 2023)
National Natural Language Processing Clinical Challenge (N2C2 2022)

PROGRAM COMMITTEE

International Conference on Computational Linguistics (COLING 2025) International Conference on
Artificial Intelligence in Education (AIED 2022)
International Conference on Language Resources and Evaluation (LREC 2022, 2024)
International Workshop on Evaluation and Comparison for NLP systems (Eval4NLP 2020-2022)
China National Conference on Computational Linguistics (CCL 2020)

REVIEWER

Association of Computational Linguistics (ACL)
International Conference Artificial Intelligence in Education (AIED)
International Conference on Language Resources and Evaluation (LREC)
Journal of Language Resources and Evaluation
Journal of Biomedical Informatics (JBI)
Journal of American Medical Informatics Association (JAMIA)
Journal of Healthcare Informatics Research
npj Digital Medicine, Nature
Scientific Reports, Nature
BMC medical informatics and decision making

TEACHING EXPERIENCES

Guest Lecturer, Department of Computer Science and Engineering, Pennsylvania State University
CSE 597, Natural Language Processing   Spring 2018 – Spring 2021

Teaching Assistant, Department of Computer Science and Engineering, Pennsylvania State University
CMPSC/DS 442, Artificial Intelligence   Fall 2018 – Spring 2021
CMPSC 221, Introduction to C++   Fall 2016 – Fall 2017

MENTORING EXPERIENCES

PhD Students

- Nourah Salem, Computational Bioscience Program, University of Colorado Anschutz, 2024 – Present

- Emma Croxford, University of Wisconsin-Madison, 2022 – Present

Master Students

- Jason Wang, California-Berkeley, 2021 – Present

Research Undergraduate Students

- Mike Chen, Pennsylvania State University, 2021 – 2022
- Lulu Liu, Pennsylvania State University, 2021 – 2022
- Steven Fontanella, Pennsylvania State University, 2019
- Andrew Warner, Pennsylvania State University, 2017 – 2018
- CMPSC 442 Honors Group Student Cohort, Pennsylvania State University, 2017 – 2019

INVITED TALKS

- " Large Language Models and Medical Knowledge Grounding for Diagnosis", Northwestern Medicine Healthcare AI Forum, 12/2023.
- " Augmented Intelligence for Healthcare: How Can Large Language Models Help Physicians at the Bedside?", PowerTalks Seminar Series, University of Alabama at Birmingham 12/2023.
- " Leverage A Medical Knowledge Graph into Large Language Models for Diagnosis Prediction", American Medical Informatics Association Natural Language Processing (NLP) Pre-symposium, 11/2023.
- " Augmented Intelligence for Healthcare: How Can NLP Help Physicians at the Bedside?", Department of Biomedical Informatics Seminar, University of Colorado, 04/2023.
- "Clinical Natural Language Processing Benchmarks for Augmented Intelligence in Healthcare", American Medical Informatics Association (AMIA) 2023 Informatics Summit, Seattle, 03/2023.
- "Text Understanding and Generation for Knowledge-Driven Decision-Making", University of California, Merced. 03/2023.
- "Text Understanding and Generation for Knowledge-Driven Decision-Making", University of Massachusetts Dartmouth. 01/2023.
- "Hierarchical Annotation for Building a Suite of Clinical Natural Language Processing Tasks", Monthly NLP Working Group Meeting, Observational Health Data Sciences and Informatics (OHDSI). 12/2022.
- "Augmented Intelligence in Healthcare: How can NLP help physicians at the bedside? " Computation and Informatics in Biology and Medicine (CIBM) Seminar, UW-Madison. 11/2022.
- "DR. BENCH: Diagnostic Reasoning Benchmark for Clinical Natural Language Processing", Department of Medicine Research Day Annual Seminar, UW-Madison, 10/2022.
- "Language Tasks for Clinical Natural Language Processing: Where We Are and What's Next. " NLP Summit, 10/2022.
- "Augmented Intelligence in Healthcare: How can NLP help physicians at the bedside?" Department of Computer Science, Loyola University Chicago, 09/2022
- "Augmented Intelligence in Healthcare: How can NLP help physicians at the bedside?" Department of Computer Science and Engineering, Penn State, 09/2022
- "Hierarchical Annotation For Building A Suite Of Clinical Natural Language Processing Tasks: Progress Note Understanding." LREC 2022. 05/2022.
- "How I Teach Machine to Understand Human Language. " School of Engineering Design, Technology and Professional Program, Penn State. 03/2022.

- "Progress Note Understanding: A Suite of Tasks for Clinical Natural Language Modeling. " UW-Madison ML+X. 03/2022

- "Medical Progress Note Annotation for Building Clinical Natural Language Processing Models". UW-Madison Postdoctoral Research Symposium. 09/2021

- "Pyreval: An automated method for summary content analysis." CLIEDE 09/2017.