

John Wieting

Contact Information

E-mail: john.wieting494@gmail.com

<https://www.cs.cmu.edu/~jwieting/>

<https://github.com/jwieting>

Last updated: July 2026

Work Experience

- Google Research / Google Brain / Google DeepMind **2020–Present**
Senior Research Scientist, Google DeepMind (2024–Present); *Research Scientist*, Google DeepMind (2023–2024), Google Brain (2022–2023), Google Research (2020–2022)
 - Own AI-generated-content (AGC) detection and filtering for Gemini pre-training data: designed the detection models, filtering and precision-estimation methodology, and gold-standard evaluations, covering English and over 100 languages; deployed from Gemini 1.5 through the latest Gemini releases, removing roughly 10% of the pre-training corpus with neutral impact on general model capabilities and reduced training time. Also initiated and guided the removal of machine-translated text from multilingual pre-training data.
 - Contributor to the Gemini, Gemma, and PaLM 2 model families through pre-training data quality work; named author on the Gemma 3, Gemma 4, PaLM 2, and Gemini technical reports.
 - Tech lead for the next-generation document quality model used to filter Gemini pre-training data, setting and executing strategy across evaluation design, prompt engineering, human rating protocols, ablation studies, and distillation.
 - Led development of unified data curation infrastructure for Gemini pre-training web data, consolidating annotation pipelines used across multiple teams and reducing corpus processing time from months to weeks (Google Perfy Gold Award).
 - Core contributor to SignGemma, Google DeepMind’s open sign language translation models: built the synthetic data (paraphrasing and translation) pipeline and its quality evaluations, enabling data compliance and significant translation performance gains, and provided the inference foundation for the self-training and reinforcement learning setup.
 - With a team of 4, built models to reliably detect automatically generated content, launched in Google Search to demote AGC websites ([Search Blog](#)), in YouTube to remove AI-generated comments, and in Google Reviews to remove AI-generated reviews.
 - Built AutoAIS (Automatic Attribution to Identified Sources) used in Google SGE ([Search Blog](#)), including the multilingual and cross-lingual models, used for automatic evaluation, as an RL reward signal for the user-facing response generator, and for filtering distillation data.
- Scientific Adviser at QuillBot **2018–2020**
Advised the founders on NLP technology from shortly after the company’s founding, through two funding rounds, to its eventual acquisition.

Education

- Ph.D. in Language Technologies, School of Computer Science **2017–2020**
Carnegie Mellon University
Ph.D. Thesis: Learning and Applications of Paraphrastic Representations for Natural Language
Advisers: Prof. Graham Neubig, Prof. Taylor Berg-Kirkpatrick
- Ph.D. student in Computer Science (transferred to CMU) **2015–2017**
Toyota Technological Institute at the University of Chicago
Adviser: Prof. Kevin Gimpel
- Ph.D. student in Computer Science (transferred to TTIC) **2014–2015**
University of Illinois, Urbana-Champaign
Adviser: Prof. Dan Roth
- M.S. in Computer Science **2014**
University of Illinois, Urbana-Champaign
M.S. Thesis: Lexical Entailment
Adviser: Prof. Dan Roth

Selected Coursework: Statistical Learning Theory, Machine Learning Theory, Nonparametric Bayesian Methods, Nonlinear Programming

- B.S. Mathematics, B.S. Chemistry with Honors
University of Wisconsin, Madison
Honors Program in Letters and Science

2011

Publications

1. [POLARIS: Guiding Small Models to Write Long Stories](#)
R. Rajendhran, J. Russell, M. Iyyer, **J. Wieting**
Preprint on **arXiv**, 2026
2. [StoryScope: Investigating Idiosyncrasies in AI Fiction](#)
J. Russell, R. Rajendhran, C. Pham, M. Iyyer, **J. Wieting**
Preprint on **arXiv**, 2026
3. [Gemma 4 Technical Report](#)
Gemma Team: S. El Abd, ..., **J. Wieting**, et al.
Technical report, **Google DeepMind**, 2026
4. [Improving Detection of Watermarked Language Models](#)
D. Bahri, **J. Wieting**
Transactions on Machine Learning Research (TMLR), 2026
5. [A Watermark for Black-Box Language Models](#)
D. Bahri, **J. Wieting**, D. Alon, D. Metzler
Transactions on Machine Learning Research (TMLR), 2026
6. [Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities](#)
G. Comanici, ..., **J. Wieting**, et al.
Technical report, **Google DeepMind**, 2025
7. [Gemma 3 Technical Report](#)
Gemma Team: A. Kamath, ..., **J. Wieting**, et al.
Technical report, **Google DeepMind**, 2025
8. [A Gold Standard Dataset for the Reviewer Assignment Problem](#)
I. Stelmakh, **J. Wieting**, Y. Xi, G. Neubig, N. Shah
Transactions on Machine Learning Research (TMLR), 2025
9. [PostMark: A Robust Blackbox Watermark for Large Language Models](#)
Y. Chang, K. Krishna, A. Houmansadr, **J. Wieting**, M. Iyyer
Proceedings of **EMNLP**, 2024
10. [Multiple References with Meaningful Variations Improve Literary Machine Translation](#)
S. Wu, **J. Wieting**, D. Smith
Preprint on **arXiv**, 2024
11. [Gemini 1.5: Unlocking Multimodal Understanding across Millions of Tokens of Context](#)
Gemini Team: M. Reid, ..., **J. Wieting**, et al.
Technical report, **Google DeepMind**, 2024
12. [Leveraging LLMs for Synthesizing Training Data across many Languages in Multilingual Dense Retrieval](#)
N. Thakur, J. Ni, G. Ábrego, **J. Wieting**, J. Lin, D. Cer
Proceedings of **NAACL**, 2024
13. [Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval is an Effective Defense](#)
K. Krishna, Y. Song, M. Karpinska, **J. Wieting***, M. Iyyer*
Proceedings of **NeurIPS**, 2023

14. [Gemini: A Family of Highly Capable Multimodal Models](#)
Gemini Team: R. Anil, ..., **J. Wieting**, et al.
Technical report, **Google DeepMind**, 2023
15. [FIAT: Fusing Learning Paradigms with Instruction-Accelerated Tuning](#)
X. Wang, **J. Wieting**, J. Clark
Preprint on **arXiv**, 2023
16. [QA is the New KR: Question-Answer Pairs as Knowledge Bases](#)
W. Cohen, W. Chen, M. de Jong, N. Gupta, A. Presta, P. Verga, **J. Wieting**
Proceedings of **AAAI**, 2023
17. [Evaluating and Modeling Attribution for Cross-Lingual Question Answering](#)
B. Muller, **J. Wieting**, J. Clark, T. Kwiatkowski, S. Ruder, L. Baldini-Soares, R. Aharoni, J. Herzig, X. Wang
Proceedings of **EMNLP**, 2023
18. [XTREME-UP: A User-Centric Scarce-data Benchmark for Under-represented Languages](#)
S. Ruder, ..., **J. Wieting**, et al.
Proceedings of **Findings of EMNLP**, 2023
19. [Evaluating Large Language Models on Controlled Generation Tasks](#)
J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, **J. Wieting**, N. Peng, X. Ma
Proceedings of **EMNLP**, 2023
20. [PaLM 2 Technical Report](#)
R. Anil, ..., **J. Wieting**, et al.
Technical report, **Google**, 2023
21. [Beyond Contrastive Learning: A Variational Generative Model for Multilingual Retrieval](#)
J. Wieting, J. Clark, W. Cohen, G. Neubig, T. Berg-Kirkpatrick
Proceedings of **ACL**, 2023
22. [Augmenting Pre-trained Language Models with QA-memory for Open-domain Question Answering](#)
W. Chen, P. Verga, M. de Jong, **J. Wieting**, W. Cohen
Proceedings of **EACL**, 2023
23. [Exploring Document-Level Literary Machine Translation with Parallel Paragraphs from World Literature](#)
K. Thai, M. Karpinska, K. Krishna, B. Ray, M. Inghilleri, **J. Wieting**, M. Iyyer
Proceedings of **EMNLP**, 2022 (Oral presentation)
24. [RankGen: Improving Text Generation with Large Ranking Models](#)
K. Krishna, Y. Chang, **J. Wieting**, M. Iyyer
Proceedings of **EMNLP**, 2022 (Oral presentation)
25. [Faithful to the Document or to the World? Mitigating Hallucinations via Entity-Linked Knowledge in Abstractive Summarization](#)
Y. Dong, **J. Wieting**, P. Verga
Proceedings of **Findings of EMNLP**, 2022
26. [On the Ingredients of an Effective Zero-Shot Semantic Parser](#)
P. Yin, **J. Wieting**, A. Sil, G. Neubig
Proceedings of **ACL**, 2022
27. [Paraphrastic Representations at Scale](#)
J. Wieting, K. Gimpel, G. Neubig, T. Berg-Kirkpatrick
Proceedings of **System Demonstrations of EMNLP**, 2022
28. [Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation](#)
J. Clark, D. Garrette, I. Turc, **J. Wieting**

TACL, 2022 (Oral presentation)

29. [Improving the Diversity of Unsupervised Paraphrasing with Embedding Outputs](#)
M. Jegadeesan, S. Kumar, **J. Wieting**, Y. Tsvetkov
Proceedings of **Workshop on Multilingual Representation Learning**, 2021
30. [On Learning Text Style Transfer with Direct Rewards](#)
Y. Liu, G. Neubig, **J. Wieting**
Proceedings of **NAACL**, 2021
31. [Reformulating Unsupervised Style Transfer as Paraphrase Generation](#)
K. Krishna, **J. Wieting**, M. Iyyer
Proceedings of **EMNLP**, 2020
32. [Improving Candidate Generation for Low-Resource Cross-Lingual Entity Linking](#)
S. Zhou, S. Rijhwani, **J. Wieting**, J. Carbonell, G. Neubig
TACL, 2020
33. [A Bilingual Generative Transformer for Semantic Sentence Embedding](#)
J. Wieting, G. Neubig, T. Berg-Kirkpatrick
Proceedings of **EMNLP**, 2020
34. [Beyond BLEU: Training Neural Machine Translation with Semantic Similarity](#)
J. Wieting, T. Berg-Kirkpatrick, K. Gimpel, G. Neubig
Proceedings of **ACL**, 2019 (Oral presentation)
35. [Simple and Effective Paraphrastic Similarity from Parallel Translations](#)
J. Wieting, K. Gimpel, G. Neubig, T. Berg-Kirkpatrick
Proceedings of **ACL**, 2019
36. [compare-mt: A Tool for Holistic Comparison of Language Generation Systems](#)
G. Neubig, Z. Dou, J. Hu, P. Michel, D. Pruthi, X. Wang, **J. Wieting**
Preprint on **arXiv**, 2019
37. [No training required: Exploring Random Encoders for Sentence Classification](#)
J. Wieting, D. Kiela
Proceedings of **ICLR**, 2019
38. [Quality Signals in Generated Stories](#)
M. Sagarkar, **J. Wieting**, L. Tu, K. Gimpel
Proceedings of ***SEM**, 2018
39. [CogCompNLP: Your Swiss Army Knife for NLP](#)
D. Khashabi, M. Sammons, B. Zhou, T. Redman, C. Christodoulopoulos, V. Srikumar, N. Rizzolo, L. Ratnov, G. Luo, Q. Do, C. Tsai, S. Roy, S. Mayhew, Z. Feng, **J. Wieting**, X. Yu, Y. Song, S. Gupta, S. Upadhyay, N. Arivazhagan, Q. Ning, S. Ling, D. Roth
Proceedings of **LREC**, 2018
40. [Adversarial Example Generation with Syntactically Controlled Paraphrase Networks](#)
M. Iyyer*, **J. Wieting***, K. Gimpel, L. Zettlemoyer
Proceedings of **NAACL**, 2018 (Oral presentation)
41. [ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations](#)
J. Wieting and K. Gimpel
Proceedings of **ACL**, 2018
42. [Learning Paraphrastic Sentence Embeddings from Back-Translated Bitext](#)
J. Wieting, J. Mallinson, and K. Gimpel
Proceedings of **EMNLP**, 2017
43. [Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings](#)
J. Wieting and K. Gimpel

44. [Charagram: Embedding Words and Sentences via Character n-grams](#)
J. Wieting, M. Bansal, K. Gimpel, and K. Livescu
Proceedings of **EMNLP**, 2016
45. [UMD-TTIC-UW at SemEval-2016 Task 1: Attention-Based Multi-Perspective Convolutional Neural Networks On Textual Similarity Measurement](#)
H. He, **J. Wieting**, K. Gimpel, J. Rao, and J. Lin
Proceedings of **SemEval**, 2016 (Best of **SemEval**)
46. [Towards Universal Paraphrastic Sentence Embeddings](#)
J. Wieting, M. Bansal, K. Gimpel, and K. Livescu
Proceedings of **ICLR**, 2016 (Oral presentation)
47. [From Paraphrase Database to Compositional Paraphrase Model and Back](#)
J. Wieting, M. Bansal, K. Gimpel, K. Livescu, and D. Roth
TACL, 2015 (Oral presentation at **EMNLP**, 2015)
48. [Clustering With Side Information: From a Probabilistic Model to a Deterministic Algorithm](#)
D. Khashabi, **J. Wieting**, J. Liu, and F. Liang
Preprint on **arXiv**, 2015
49. [Tiered Clustering to Improve Lexical Entailment](#)
J. Wieting
Preprint on **arXiv**, 2014
50. [Illinois Cognitive Computation Group UI-CCG TAC 2013 Entity Linking and Slot Filler Validation Systems](#)
X. Cheng, B. Chen, R. Samdani, K-W. Chang, Z. Fei, M. Sammons, **J. Wieting**, S. Roy, C. Wang, and D. Roth
Proceedings of **TAC**, 2013 (Presented the Slot Filler Validation System at NIST, November 2013)

Software and Data

- [ParaNMT-50M](#)
50 million English paraphrase pairs produced by back-translating the Czech side of CzEng, released with trained sentence embedding models (from **ACL** 2018).
- [Paraphrastic representations at scale](#)
Trained paraphrastic sentence embedding models for English and 90+ languages, with a small interface for embedding text (from **EMNLP** 2022).
- [compare-mt](#)
Tool for holistic comparison and analysis of the outputs of language generation systems.

Internships

- Facebook AI Research - New York **May–August 2018**
Advisers: Douwe Kiela, Kyunghyun Cho, Jason Weston
- Microsoft Research - Redmond **June–September 2016**
Advisers: Michel Galley, Jianfeng Gao, Bill Dolan, Chris Brockett
- Toyota Technological Institute - Chicago **May–August 2015**
Advisers: Prof. Mohit Bansal, Prof. Kevin Gimpel, Prof. Karen Livescu
- Toyota Technological Institute - Chicago **May–August 2014**
Advisers: Prof. Mohit Bansal, Prof. Kevin Gimpel, Prof. Karen Livescu

Honors and Awards

- Google Perfy Award, Gold (Machine Learning category), for data curation infrastructure for Gemini pre-training, 2025
- Google Research PA Tech Impact Award, for contributions to large language model training, Q1 2023
- Outstanding Reviewer, ACL, 2018
- Best of SemEval (SemEval-2016 Task 1: Semantic Textual Similarity), 2016
- State Farm Doctoral Fellowship, 2014

	• Outstanding CS Teaching Assistant, 2013 • UW-Madison Undergraduate Summer Chemistry Research Fellowship, 2008
Grants	• Google DeepMind External Research Grant Proposed, secured, and sponsor a \$30,000 unrestricted grant funding research on long-form text generation in Prof. Mohit Iyyer's group, 2025 • NSF XSEDE Allocation Co-authored with Taylor Berg-Kirkpatrick for 3,500 GPU hours, 2020 • Microsoft Azure for Research Award Co-authored with Kevin Gimpel. Award was for \$20,000, 2016 • Argonne Leadership Computing Facility Director's Discretionary Allocation Award Co-authored with Kevin Gimpel. Award was for 10,000 GPU hours, 2016
Student Mentorship	• Ramon Sanabria, Intern at Google DeepMind, (2023) • Nandan Thakur, Intern at Google Research, (2023) • Benjamin Muller, Intern at Google Brain (Primary Host), (2022) • Kalpesh Krishna, Student Researcher at Google Brain (Primary Host), (2021–2022) • Yue Dong, Intern at Google Research, (2021) • Yixin Liu, Masters student at CMU, now Ph.D. student at Yale, (2019–2021)
Invited Talks	• Georgia Institute of Technology, CS 4650. Natural Language Understanding, (2024) • Google, NLX Invited Talk, (2023) • Google, Lux Leads Meeting, (2023) • Ohio State University, (2020) • University of Massachusetts - Amherst, (2020) • Allen Institute of Artificial Intelligence, (2020) • IBM, (2020) • Midwest Speech and Language Days at University of Chicago, (2017) • Midwest Speech and Language Days at University of Chicago, (2015) • CCICADA Research Retreat at Carnegie Mellon University, (2015)
Teaching	• Teaching Assistant for Machine Translation and Sequence-to-sequence Models (CMU CS11-731), Fall 2018 • Teaching Assistant for Advanced Natural Language Processing (TTIC 31210), Spring 2017 • Teaching Assistant for Machine Learning and Natural Language (UIUC CS546), Spring 2013 • Teaching Assistant for Programming Languages and Compilers (UIUC CS421), Summer 2012 • Teaching Assistant for Intro to Computer Science (for Majors) (UIUC CS125), Fall 2012 <i>Excellent Teacher List</i> • Teaching Assistant for Intro to Computer Science (for Majors) (UIUC CS125), Spring 2012 <i>Excellent Teacher List, Outstanding Teacher List (top 10%)</i> • Teaching Assistant for Intro to Computer Science (for Majors) (UIUC CS125), Fall 2011 <i>Excellent Teacher List</i>
Professional Service	• ACL 2020 Paper Matching System Designed and trained the abstract similarity model in a system used to assign reviewers to papers for ACL 2020 based on my earlier research. I then helped build the system with my adviser, Graham Neubig. • Area Chair – <i>NeurIPS</i> 2023, 2024, 2025 – <i>ACL</i> 2020, 2024 – <i>NAACL</i> 2021, 2024 – <i>EMNLP</i> 2024 – <i>ICML</i> 2025 – <i>AAAI</i> 2022, 2023, 2026 (Senior Program Committee, 2027) – <i>ARR</i> (ongoing) – <i>EACL</i> 2024

- **Reviewing**

- *ACL* 2018 (Outstanding Reviewer), 2019, 2021
- *EMNLP* 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023
- *NAACL* 2018, 2019
- *ICLR* 2018, 2019, 2020, 2024
- *NeurIPS* 2018, 2019, 2020
- *COLM* 2024
- *ICML* 2019, 2021
- *AAAI* 2020
- *COLING* 2016, 2018
- *CoNLL* 2017, 2019, 2020
- **SEM* 2017, 2019, 2020
- *LREC* 2018, 2020
- *NAACL Student Research Workshop* 2018, 2019, 2021
- *NeuralGen: Workshop on Methods for Optimizing and Evaluating Neural Language Generation* 2019
- *Deep Learning Inside Out (DeeLIO) Workshop* 2021
- *NeurIPS Reproducibility Challenge* 2019
- *ML Reproducibility Challenge* 2020
- *Transactions on Machine Learning Research (TMLR)*
- *Journal of Natural Language Engineering*
- *ACM Transactions on Asian and Low-Resource Language Information Processing*

- **Graduate student representative on General Education Self-Study Committee**

Committee was tasked with both analyzing the state of courses satisfying general education requirements and identifying exemplar general education courses at the University of Illinois.