William Merrill

https://lambdaviking.com/
willm[æt]nyu.edu

Last updated May 2, 2024

Research Interests

Broad Applications of the following to analyzing language models, NLP, and linguistics:

- Formal languages and automata
- Computational complexity; especially circuits
- Formal semantics

Specific Two key problems I have worked on are:

- Expressive power and inductive biases of neural nets for implementing algorithms, representing the syntax and semantics of natural language, and reasoning
- The theory of learning linguistic semantics from text corpora

Experience

Allen Institute for AI	2023	**Research Intern**, AllenNLP
Google Research	2022	**Student Researcher**, Speech & Lang. Algorithms
New York University	2021–	**Ph.D.**, Center for Data Science
Allen Institute for AI	2019–21	**PYI** (predoc. researcher), AllenNLP
Yale University	2015–19	**B.Sc.** with distinction in Computer Science **B.A.** with distinction in Linguistics Thesis: Sequential neural networks as automata *Cum laude; note of excellence on thesis*
Google	2018	**Software Engineering Intern** *“Exceeds expectations" rating; return offer*
Boston College	2017	**Research Intern**, Language Learning Lab
New York University	2013–15	**Research Intern**, Morphology Lab

Academic Group Affiliations

CapLab & **ML^2**, NYU	*Tal Linzen*	2021–
AllenNLP, AI2	*Noah A. Smith, Yoav Goldberg, Roy Schwartz*	2019–21
CLAY, Yale	*Robert Frank, Dana Angluin*	2016–19
L^3, Boston College	*Joshua Hartshorne, Sven Dietz*	2017
MorphLab, NYU	*Alec Marantz, Phoebe Gaston*	2013–15

Archival Publications

[1] **W. Merrill**, J. Petty, and A. Sabharwal. The Illusion of State in State-Space Models. *ICML*. July 2024.

[2] M. Zhang, O. Press, **W. Merrill**, A. Liu, and N. A. Smith. How Language Model Hallucinations Can Snowball. *ICML*. July 2024.

[3] **W. Merrill** and A. Sabharwal. The Expressive Power of Transformers with Chain of Thought. *ICLR*. May 2024.

[4] **W. Merrill** and A. Sabharwal. A Logic for Expressing Log-Precision Transformers. *NeurIPS*. Dec. 2023.

[5] **W. Merrill**. Formal Languages and the NLP Black Box. *Developments in Language Theory*. Ed. by F. Drewes and M. Volkov. Cham: Springer Nature Switzerland, June 2023.

[6] **W. Merrill** and A. Sabharwal. The Parallelism Tradeoff: Limitations of Log-Precision Transformers. *TACL* (June 2023).

[7] **W. Merrill**, N. Tsilivis, and A. Shukla. A Tale of Two Circuits: Grokking as Competition of Sparse and Dense Subnetworks. *ICLR Workshop on Mathematical and Empirical Understanding of Foundation Models*. May 2023.

[8] Z. Wu, **W. Merrill**, H. Peng, I. Beltagy, and N. A. Smith. Transparency Helps Reveal When Language Models Learn Meaning. *TACL* (2023).

[9] **W. Merrill**, A. Warstadt, and T. Linzen. Entailment Semantics Can Be Extracted from an Ideal Language Model. *CoNLL*. Abu Dhabi, United Arab Emirates (Hybrid), Dec. 2022.

[10] **W. Merrill**, A. Sabharwal, and N. A. Smith. Saturated Transformers are Constant-Depth Threshold Circuits. *TACL* (Aug. 2022).

[11] S. Subramanian, **W. Merrill**, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach. ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension. *ACL*. Dublin, Ireland, May 2022.

[12] M. Gardner, **W. Merrill**, J. Dodge, M. Peters, A. Ross, S. Singh, and N. A. Smith. Competency Problems: On Finding and Removing Artifacts in Language Data. *EMNLP*. Online and Punta Cana, Dominican Republic, Nov. 2021.

[13] **W. Merrill**, V. Ramanujan, Y. Goldberg, R. Schwartz, and N. A. Smith. Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent. *EMNLP*. Online and Punta Cana, Dominican Republic, Nov. 2021.

[14] **W. Merrill**, Y. Goldberg, R. Schwartz, and N. A. Smith. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *TACL* (Sept. 2021).

[15] **W. Merrill**, G. Weiss, Y. Goldberg, R. Schwartz, N. A. Smith, and E. Yahav. A Formal Hierarchy of RNN Architectures. *ACL*. Online, July 2020.

[16] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, **W. Merrill**, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier. CORD-19: The COVID-19 Open Research Dataset. *ACL Workshop on NLP for COVID-19*. Online, July 2020.

[17] **W. Merrill**. Sequential Neural Networks as Automata. *ACL Workshop on Deep Learning and Formal Languages*. Florence, Aug. 2019.

[18] **W. Merrill**, L. Khazan, N. Amsel, Y. Hao, S. Mendelsohn, and R. Frank. Finding Hierarchical Structure in Neural Stacks Using Unsupervised Parsing. *ACL Workshop BlackboxNLP*. Florence, Italy, Aug. 2019.

[19] **W. Merrill**, G. Stark, and R. Frank. Detecting Syntactic Change Using a Neural Part-of-Speech Tagger. *ACL Workshop on Computational Approaches to Historical Language Change*. Florence, Italy, Aug. 2019.

[20] Y. Hao, **W. Merrill**, D. Angluin, R. Frank, N. Amsel, A. Benz, and S. Mendelsohn. Context-Free Transductions with Neural Stacks. English. *EMNLP Workshop BlackboxNLP*. Brussels, Belgium, Nov. 2018.

[21] J. Kasai, R. Frank, P. Xu, **W. Merrill**, and O. Rambow. End-to-End Graph-Based TAG Parsing with Neural Networks. *NAACL*. 2018.

Non-Archival Publications

[22] J. Pfau, **W. Merrill**, and S. Bowman. *Let's Think Dot by Dot: Hidden Computation in Transformer Language Models*. Apr. 2024.

[23] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Jha, H. Ivison, I. Magnusson, Y. Wang, S. Arora, D. Atkinson, R. Authur, K. R. Chandu, A. Cohan, J. Dumas, Y. Elazar, Y. Gu, J. Hessel, T. Khot, **W. Merrill**, J. Morrison, N. Muennighoff, A. Naik, C. Nam, M. E. Peters, V. Pyatkin, A. Ravichander, D. Schwenk, S. Shah, W. Smith, E. Strubell, N. Subramani, M. Wortsman, P. Dasigi, N. Lambert, K. Richardson, L. Zettlemoyer, J. Dodge, K. Lo, L. Soldaini, N. A. Smith, and H. Hajishirzi. *OLMo: Accelerating the Science of Language Models*. Feb. 2024.

[24] **W. Merrill**, Z. Wu, N. Naka, Y. Kim, and T. Linzen. *Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment*. Feb. 2024.

[25] **W. Merrill** and N. Tsilivis. *Extracting Finite Automata from RNNs Using State Merging*. Jan. 2022.

[26] **W. Merrill**. *Formal Language Theory Meets Modern NLP*. Feb. 2021.

[27] **W. Merrill**. *On the Linguistic Capacity of Real-Time Counter Automata*. Sept. 2020.

[28] **W. Merrill**. A Semantics of Subordinate Clauses Using Delayed Evaluation. *Toronto Undergraduate Linguistics Conference* (2018).

Press Coverage

[1] **Quanta Magazine**. How Chain-of-Thought Reasoning Helps Neural Networks Compute. March 2024.

[2] **NYU CDS Blog**. Language Models Provide Insight into Linguistic Redundancy. March 2024.

[3] **Washington Post**. Honestly, I Love When AI Hallucinates. Dec. 2023.

[4] **NYU CDS Blog**. The Logic of Transformers: William Merrill's Step Towards Understanding Large Language Models' Limits and Hallucinations. Oct 2023.

[5] **NYU CDS Blog**. Can Language Models Learn Meaning Just By Observing Text?. Oct 2022.

Invited Talks

[1] **Brown**, LUNAR Lab, 2024
Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment

[2] **NYU**, Guest Speaker (Natural Language Understanding), 2024
The Parallelism Tradeoff: Limitations of Log-Precision Transformers

[3] **FLaNN**, 2024
The Illusion of State in State-Space Models

[4] **Flatiron Institute**, Transformer Theory Seminar, 2024
Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment

[5] **Two Sigma**, Fellowship Finalist Reception, 2024
Provable Reasoning Limitations of Transformer Language Models

[6] **Flatiron Institute**, Transformer Theory Seminar, 2023
The Parallelism Tradeoff: Limitations of Log-Precision Transformers

[7] **Limitations of LMs Workshop**, Bielefeld University, 2023
The Parallelism Tradeoff: Limitations of Log-Precision Transformers

[8] **AI2**, AllenNLP, 2023
Benchmarking Whether LMs Copy from Their Pretraining Data

[9] **MIT CSAIL**, Lingo Group, 2023
The Parallelism Tradeoff: Limitations of Log-Precision Transformers

[10] **Institut Jean Nicod**, Linguae Seminar, 2023
Entailment Semantics Can Be Extracted from an Ideal Language Model

[11] **ICGI**, Invited Speaker at Conference, 2023
Formal Languages and Neural Models for Learning on Sequences

[12] **Developments in Language Theory**, Invited Speaker at Conference, 2023
Formal Languages and the NLP Black Box

[13] **NYC Philosophy of Language Workshop**, Invited Speaker, 2023
Entailment Semantics Can Be Extracted from an Ideal Language Model

[14] **NYU**, Depth Qualifying Exam, 2023
Transformer Reasoning Through the Lens of Circuit Complexity

[15] **NYU**, Guest Speaker (Comp. Ling. & Cognitive Science), 2023
Entailment Semantics Can Be Extracted From an Ideal Language Model

[16] **EMNLP**, TACL Track, 2022
Saturated Transformers are Constant-Depth Threshold Circuits

[17] **CoNLL**, 2022
Entailment Semantics Can Be Extracted From an Ideal Language Model

[18] **Microsoft Research**, New York, 2022
The Parallelism Tradeoff: Insights on the Power and Limitations of Transformers Using Circuit Complexity

[19] **Umeå University**, Foundations of Language Processing, 2022
Entailment Semantics Can Be Extracted from an Ideal Language Model

[20] **ArthurAI**, Journal Club, 2022
Entailment Semantics Can Be Extracted from an Ideal Language Model

[21] **FLaNN Discord**, Weekly Seminar, 2022
Saturated Transformers are Constant-Depth Threshold Circuits

[22] **Umeå University**, Foundations of Language Processing, 2022
Saturated Transformers are Constant-Depth Threshold Circuits

[23] **MILA**, ML for Code Seminar, 2022
Saturated Transformers are Constant-Depth Threshold Circuits

[24] **MIT**, CompLang Seminar, 2022
Language Models Have Implicit Entailment Semantics

[25] **NYU**, Semantics Seminar, 2022
Distributional Learnability of Entailment

[26] **Google**, Speech and Language Algorithms, 2022
Neural Networks as Automata

[27] **ArthurAI**, Journal Club, 2021
Competency Problems: On Finding and Removing Artifacts in Language Data

[28] **EMNLP**, ML Track, 2021
Competency Problems: On Finding and Removing Artifacts in Language Data

[29] **EMNLP**, ML Track, 2021
Parameter Norm Growth During Transformer Training: Inductive Bias From Gradient Descent

[30] **AI2**, All Hands, 2021
Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?

[31] **UW**, Noah's ARK, 2020
Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?

[32] **EMNLP**, Blackbox NLP, 2018
Context-Free Transductions with Neural Stacks

[33] **Packer Collegiate Institute**, Science Research Symposium, 2018
Neural networks, L2 Acquisition, and the Voynich

[34] **CodeHaven**, 2018
Programming, Language, and the Book of Thoth

[35] **UToronto**, TULCon, 2018
A Semantics of Subordinate Clauses Using Delayed Evaluation

Poster Presentations

[1] **NeurIPS**, M3L Workshop, 2024
The Expressive Power of Transformers with Chain of Thought

[2] **NeurIPS**, 2024
A Logic for Expressing Log-Precision Transformers

[3] **Philosophy of Deep Learning Workshop**, NYU, 2023
Entailment Semantics Can Be Extracted from an Ideal Language Model

[4] **EMNLP**, ML Track, 2021
Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent

[5] **EMNLP**, ML Track, 2021
Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?

[6] **ACL**, Deep Learning and Formal Languages, 2019
Sequential Neural Networks as Automata

[7] **ACL**, Blackbox NLP, 2019
Finding Hierarchical Structure in Neural Stacks Using Unsupervised Parsing

Teaching Experience

University Level

[1] **Lead TA** for *Natural Language Processing*, Tal Linzen (NYU, Fall 2022)

[2] **TA** for introductory NLP (NYC AI School, Spring 2022)

[3] **TA** for *Artificial Intelligence*, Dragomir Radev (Yale, Spring 2019)

[4] **TA** for *Natural Language Processing*, Dragomir Radev (Yale, Fall 2018)

[5] **TA** for *Artificial Intelligence*, Dragomir Radev (Yale, Spring 2017)

High-School Level and Below

[6] Instructor for CodeHaven (Yale, 2016-2018)

[7] Designed and taught *Viking Runes* (Yale Splash, Spring 2017)

[8] Taught *The Politics of Skyrim* (Yale Splash, Spring 2016)

[9] Designed and taught *DECLASSIFIED: The History of Codebreaking* (Yale Splash, Fall 2016)

Service

Reviewing

ICLR	Oct 2023	*Conference*	3 reviews (+1 emergency)
M3L	Oct 2023	*Workshop*	3 reviews
GenBench	Sept 2023	*Workshop*	3 reviews
NeurIPS	July 2023	*Conference*	1 emergency review
JMLR	June 2023	*Journal*	1 review
ACL SRW	May 2023	*Workshop*	2 reviews
ICGI	April 2023	*Conference*	2 reviews
ACL	Feb 2023	*Conference*	1 review
Proc. of Royal Society A	Jan 2023	*Journal*	1 review
ARR	Nov 2022	*Conference*	1 review
Inverse Scaling Prize	Sept 2022	*Competition*	7 reviews
TheoretiCS	July 2022	*Journal*	1 review
ARR	April 2022	*Conference*	1 review
ARR	Jan 2022	*Conference*	2 review
ARR	Dec 2021	*Conference*	3 reviews
ARR	Nov 2021	*Conference*	1 review
CL	2021	*Journal*	1 review
ACL	2021	*Conference*	6 reviews
EACL	2021	*Conference*	4 reviews
EMNLP	2020	*Conference*	2 reviews
Neural Networks	2020	*Journal*	1 review

Session Chairing

ICGI	July 2023
DLT	June 2023

Other

[1] **NYC AI School**, organizer (2024)

[2] **ML2 Seminar**, organizer (2024)

[3] **CAP Lab Website**, maintainer (2023)

[4] **FLaNN Discord**, moderator, scheduled and hosted talks (2022)

[5] **NYC AI School**, volunteer instructor (2022)

[6] **AllenNLP Hackathon**, technical support (2021)

[7] **AllenNLP Tutorial**, chapter author (2020)

[8] **Yale Tangut Language Workshop**, videographer and technical support (2018)

[9] **Yale Kitan Language Workshop**, videographer and technical support (2016)

[10] **CodeHaven**, student volunteer (2016–18)

[11] **Splash at Yale**, volunteer instructor (2016–17)

Selected Public Software

[1] **AllenNLP**: Open-source NLP framework (contributor)

[2] **The Book of Thoth**: Puzzle game with compositional spell casting in Middle Egyptian hieroglyphs

[3] **DraftNet**: Dota 2 drafting using neural networks

[4] **Voynich2Vec**: Word embedding analysis of the Voynich manuscript

[5] **StackNN**: Differentiable stacks, queues, and dequeues in PyTorch

Blog Posts

Research Content

[1] *A Formal Hierarchy of RNN Architectures* (2020)

[2] *Theory of Saturated Neural Networks* (2019)

[3] *The State of Interpretability in NLP* (2019, outdated!)

[4] *Word2vec Analysis of the Voynich Manuscript* (2018)

[5] *Review: Learning to Transduce with Unbounded Memory* (2018)

[6] *Capsule Networks for NLP* (2018)

Translations

[7] *The Wanderer* (Old English → English)

[8] *After Ragnarok* (Old Norse → English)

[9] *The Saga of Mary* (Old Norse → English)

Awards and Grants

[1] First annual **Angluin Invited Tutorial Speaker** (ICGI 2023)

[2] NSF **Graduate Student Research Fellowship** (2022)

[3] **Student Travel Grant** to attend DELFOL workshop at ACL from Naver Labs (2019)

[4] **Mellon Grant** for senior thesis from Benjamin Franklin College at Yale University(2019)

[5] **Grace Hopper Prize** for computer science finalist (2017)

[6] Yale College **freshman rap battle champion** (2016)

[7] **Rising Scientist Award** presented by the Child Mind Institute (2015)

[8] **Study of American History Award** from the Society of Mayflower Descendants (2013)

[9] National Latin Exam ***cum honore maximo egregio*** (2010)

Selected Coursework

Theoretical Computer Science and Formal Languages

[1] *Inference and Representation* (NYU, 2022)

[2] *Foundations of Machine Learning* (NYU, 2022)

[3] *Computational Complexity Theory* (Yale, 2018)

[4] *Computability and Logic* (Yale, 2017)

[5] *Design and Analysis of Algorithms* (Yale, 2017)

[6] *Computing Meanings* (Yale, 2016)

[7] *Introduction to Computer Science* (Yale, 2015)

[8] *Formal Foundations of Linguistic Theory* (Yale, 2015)

Deep Learning and Natural Language Processing

[9] *Seminar: Scaling Laws, the Bitter Lesson, and AI Research* (NYU, 2021)

[10] *Ph.D. Introduction to Data Science* (NYU, 2021)

[11] *Seminar: Selected Topics in Neural Networks* (Yale, 2019)

[12] *Seminar: Advanced Natural Language Processing* (Yale, 2018)

[13] *Computational Vision and Biological Perception* (Yale, 2018)

[14] *Neural Networks and Language* (Yale, 2018)

[15] *Deep Learning Theory and Applications* (Yale, 2018)

[16] *Natural Language Processing* (Yale, 2017)

Other Linguistics

[17] *Hybrid Grammars: Language Contact and Change* (Yale, 2019)

[18] *Phonology I* (Yale, 2018)

[19] *The Voynich Manuscript* (Yale, 2018)

[20] *Indo-European Linguistics* (Yale, 2018)

[21] *Syntax I* (Yale, 2017)

[22] *Seminar: Beowulf and the Northern Heroic Tradition* (Yale, 2017)

[23] *Medieval Latin Paleography* (Yale, 2016)

[24] *Semantics I* (Yale, 2016)

[25] *Old English* (Yale, 2015)

Other Computer Science

[26] *Big Data* (NYU, 2022)

[27] *Systems Programming Techniques and Computer Organization* (Yale, 2017)

[28] *Data Structures and Programming Techniques* (Yale, 2016)

Continuous Math

[29] *Introduction to Analysis* (Yale, 2017)

[30] *MATH 231: Vector Calculus and Linear Algebra II* (Yale, 2016)

[31] *MATH 230: Vector Calculus and Linear Algebra I* (Yale, 2015)

Reading Groups

[31] *Nonlinear Dynamical Systems* (AI2, 2021)

[32] *Deep Learning Theory* (AI2, 2020)

Languages

[1] **Modern:** English (Native), Icelandic (Intermediate)

[2] **Ancient:** Latin, Old Norse, Old English

[3] **Coding:** Python, Java, C, Rust, Haskell, PyTorch, AllenNLP, *inter alias*