Brian **Lester**

Toronto, CA

blester125.com | blester125 | blester125 | Brian Lester | Brian Lester

Education

University of Toronto — *Toronto, Canada*
Ph.D. Student working with Colin Raffel. One year was at the University of North Carolina — *2022–Present*

University of Pittsburgh — *Pittsburgh, PA*
Double major, 3.51. **Computer Science** with Honors, 3.86. **Neuroscience**, 3.19. — *2012–2016*

Shanghai American School — *Shanghai, China*
High School, Team Lead in FIRST Robotics Competition. — *2008–2012*

Selected Publications

h-index: 11

The Power of Scale for Parameter-Efficient Prompt Tuning
Brian Lester, Rami Al-Rfou, Noah Constant
EMNLP 2021, Online. **4914** citations

Finetuned Language Models Are Zero-Shot Learners
Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, **Brian Lester**, Nan Du, Andrew M Dai, Quoc V Le
ICLR 2022, Online. **4567** citations

SPoT: Better Frozen Model Adaptation through Soft Prompt Transfer
Tu Vu, **Brian Lester**, Noah Constant, Rami Al-Rfou, Daniel Matthew Cer
ACL 2022, Dublin, Ireland and Online. **313** citations

Scaling Up Models and Data with `t5x` and `seqio`
Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, **Brian Lester**, Et Al. (8/42)
JMLR 2023. **211** citations

Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation
Tu Vu, Aditya Barua, **Brian Lester**, Daniel Matthew Cer, Mohit Iyyer, Noah Constant
EMNLP 2022, Abu Dhabi, UAE and Online. **71** citations

An Effective Label Noise Model for DNN Text Classification
Ishan Jindal, Daniel Pressel, **Brian Lester**, Matthew Nokleby
NAACL 2019, Minneapolis, Minnesota. **55** citations

A Library for Rapid Modeling, Experimentation and Development of Deep Learning Algorithms targeting NLP
Daniel Pressel, Sagnik Ray Choudhury, **Brian Lester**, Yanjie Zhao, Matt Barta
ACL 2018; *NLP-OSS Workshop*, Melbourne, Australia. **16** citations

Git-Theta: A Git Extension for Collaborative Development of Machine Learning Models
Nikhil Kandpal*, **Brian Lester***, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu, Colin Raffel
ICML 2023, Honolulu Hawaii. **14** citations

Reducing Retraining by Recycling Parameter-Efficient Prompts
Brian Lester*, Joshua Yurtsever*, Siamak Shakeri, Noah Constant
Preprint arXiv:2208.05577, 2022. **13** citations

Realistic Evaluation of Model Merging for Compositional Generalization
Derek Tam*, Yash Kant*, **Brian Lester***, Igor Gilitschenski, Colin Raffel
arXiv preprint arXiv:2409.18314, 2024. **11** citations

Multiple Word Embeddings for Increased Diversity of Representation
Brian Lester, Daniel Pressel, Amy Hemmeter, Sagnik Ray Choudhury, Srinivas Bangalore
Preprint arXiv:2009.14394, 2020. **11** citations

Training LLMs over Neurally Compressed Text
Brian Lester, Jaehoon Lee, Alex Alemi, Jeffrey Pennington, Adam Roberts, Jascha Sohl-Dickstein, Noah Constant
TMLR 2024. **10** citations

The Common Pile v0.1: An 8TB Dataset of Public Domain and Openly Licensed Text
Nikhil Kandpal*, **Brian Lester***, Colin Raffel*, Et Al. (2/27)
NeurIPS 2025; *Datasets & Benchmarks*, San Diego, CA. **8** citations

Constrained Decoding for Computationally Efficient Named Entity Recognition Taggers
Brian Lester, Daniel Pressel, Amy Hemmeter, Sagnik Ray Choudhury, Srinivas Bangalore
EMNLP 2020; *Findings*, Online. **8** citations

iobes: Library for Span Level Processing
Brian Lester
ACL 2020; *NLP-OSS Workshop*, Online. **7** citations

Work Experience

Google DeepMind *Mountain View, California*
Senior Research Engineer *2020–Present*
Deep Learning research with a focus on Natural Language Processing, large pre-trained models, and zero-shot transfer.
- Prompt Tuning: an efficient method of controlling large frozen pre-trained language models based on T5. Matches performance of full fine-tuning using only 0.003% of the parameters. Open-sourced of our codebase and it has enabled 5 published papers, 1 product launch, and at least 3 more in-flight papers.
- Flan: Multitask training for a 137 billion parameter transformer-based decoder-only language model to create a model that is more effective at zero-shot prompting and performs better using Prompt Tuning.
- SPoT: Using multitask prompts as strong initialization for Prompt Tuning resulting in increased performance. Also used prompt similarity to estimate task similarity and to predict transferability.
- Added partial network training, lazy loading, and pre-filling of the auto-regressive cache to t5x, the open-source reimplementation of T5 in Jax. This final change reduced inference latency from 30 seconds to 2.4.

Interactions *Ann Arbor, Michigan*
Machine Learning Engineer *2018–2020*
Built production grade deep learning solutions and lead research efforts to push the boundaries of performance.
- Designed novel neural network architectures for calibrated intent detection, slot filling, and named entity linking using ConvNets, bLSTM-CRFs, ranking models, and transformer-based seq2seq models.
- Designed label space, annotation guidelines, and data collection method for NLU component of dialogue systems.
- Created a cloud-native model training platform based on declarative pipelines and kubernetes. Built a deployment platform that powers NLU for multiple production dialogue systems.
- Built efficient, batched implementations of complex neural network architectures such as Beam Search. My CRF implementation reduced training time by a factor of 10.

Trove *Ann Arbor, Michigan*
Lead Machine Learning Research Engineer *2017–2018*
Created a model training and serving platform that processed 200 million emails per day. Provided technical leadership to the ML team.
- Designed ConvNets for text classification to find sentences that contain questions. This powered a user-facing feature and was used to featurize the social graph created from email.
- Created neural ranking model was used to find coreferent mentions in the text and provide context to users.
- Used lexical features, as well as connectivity information in the email social graph, to identified bot accounts.

Visteon Corporation *Van Buren, Michigan*
Software Engineering Intern *2015*
Designed an adaptive system to minimize voice recognition errors based on ASR confidence scores. We patented this system and it is used in Mazda Cars.

Presentations

Prompt Tuning *University of Michigan*
University of North Carolina
Deep Learning *2021–2022*
An overview of my work on Prompt Tuning, as well as our work—Flan and SPoT—directly built on Prompt Tuning. The talk includes a collection of insights about the behavior of soft prompts aggregated from others' followup work.

NeurIPS Spotlight Talk on Mead-Baseline *NeurIPS OSS Workshop*
Deep Learning *December 2018*
A spotlight talk, at the Open Source Software workshop at NeurIPS 2018, about our open-source toolkit, Mead-Baseline.

Confidence and Calibration of Neural Network Models *EMU ML Conference*
Deep Learning *March 2020* [1]
An overview of techniques used to adjust model calibration, evaluation of models that have the ability to "reject" decision with low confidence, and their uses in the NLU unit of a production dialogue system.

Padding in Neural Networks for Natural Language Processing *A2D-NLP*
Natural Language Processing *February 2020*
A survey of NLP building blocks with a focus on correctness and the need for padding in complex situations as well as places it is unexpected, like max-pooling following a 1D convolution.

Optimization via NumPy and Cython *Michigan Python Meetup*
Numerical Computation *January 2020*
I use a series of optimizations for computing pairwise Manhattan distance to introduce core NumPy concepts and Cython to reduce the runtime from multiple hours to just seconds.

Input Representations of Deep Neural Networks *PyData Ann Arbor*
Deep Learning *October 2017*
Using learned character-compositional input representations to create Deep Neural Networks with an open vocbaulary.

Skills

Deep Learning	Extensive experience building novel Neural Network architectures, generally for NLP. High-performance training with Data and Model Parallelism, including multihost distributed training on TPU.
Infrastructure	Build and deploy with Kubernetes, Docker, Flux, MongoDB, Apache Nifi, Github Actions, and GitLab CI/CD. Experience training large neural networks on Google Cloud (GCP)
Toolkits	Jax, Flax, PyTorch, Tensorflow, NumPy, Pandas, Faiss, SpaCy, NLTK, Tensorflow-Datasets, Seaborn, and Matplotlib.
Languages	Python, Cython, Java, C, Javascript, C++, Elisp, and LaTeX.

[1] Canceled due to COVID-19

Service & Public Scholarship

Reviewing

2023–2025	NeurIPS
2022	IEEE Transactions on Affective Computing
2022	ARR: ACL Rolling Review
2022–2024	NAACL
2025	ACL
2025	EMNLP
2025	ICLR
2025	ICML
2020–2021	Computer Speech and Language
2019	CoNLL

Public Scholarship

2020 [1] **a^2-dlearn:** Helped organize logistics, recruit speakers, and acquire funding through sponsorships