Raphael Tang

Principal Research Scientist
Curriculum Vitae, June 2024

http://ralphtang.com
1-609-516-4789 | tang.raphael@gmail.com

Research

I seek to understand and to democratize machine learning models, both empirically and theoretically, with applications in natural language processing and computer vision. Specifically, I aim to better expose and explain the inner workings of deep learning models, as well as to reduce the amount of computation and labor needed for training and running such models. I believe that these two activities form a positive feedback loop: improved understanding accelerates the theory for democratization, and improved democratization accelerates the experimentation for understanding.

Experience

Research

Principal Research Scientist Feb 2024–Present
Comcast AI Technologies *Washington, District of Columbia*

• **GenAI strategies**: leading, prototyping, and developing generative AI strategies with annual multimillion-dollar impact in customer care, speech processing, and internal operations, publishing multiple papers [1, 2, 3, 4].

Lead Research Scientist October 2022–Feb 2024
Comcast Applied AI *Washington, District of Columbia*

• **Interpreting Stable Diffusion**: implemented and led the technical experimentation and creative vision for a visuo-linguistic analysis of Stable Diffusion, publishing a first-author paper at ACL 2023 [5] that won a *best paper award*.

• **Industrial use of large language models (LLMs)**: educated 400 employees on the fundamentals of LLMs; currently assisting with companywide guidelines for large language models, the scope covering thousands of employees.

Research Scientist August 2020–October 2022
Comcast Applied AI *Washington, District of Columbia*

• **Data- and computation-efficient speech recognition**: developed weak supervision and model acceleration methods for speech transformers, publishing a first-author paper at EMNLP 2022 Industry [6] and filing a patent [33]. Also led and implemented its practical deployment, which currently transcribes 20 million real-time voice queries per day.

• **Speculative query execution for reduced latency**: ideated and explored a latency-reducing approach that speculatively executed autocompleted natural language queries, publishing first-author papers at EMNLP 2021 [10] and ICASSP 2022 [9] and filing two related patents [30, 31].

• **Interpreting pretrained language transformers**: advised papers about interpreting pretrained transformers for natural language processing, publishing at EMNLP 2020 Findings [13] and BlackboxNLP 2021 workshop [24].

• **Research management**: started and coordinated research discussions about natural language processing among more than ten teammates, resulting in the filing of three patents [30, 31, 32].

Graduate Research Assistant January 2018–August 2020
University of Waterloo *Waterloo, Ontario*

• **Model acceleration**: proposed memory/latency reduction methods for models in natural language processing and computer vision, including being the first to accelerate pretrained transformers (BERT) [38], with publications at a NeurIPS 2018 workshop [29], an EMNLP 2019 workshop [28], ACL 2020 [15], and EACL 2021 [12].

• **Wake-word detection system**: collaborated with a Mozilla team to develop a wake-word detection system for Firefox Voice, publishing a first-author paper at an EMNLP workshop [25].

• **Experimental validity**: constructed an unbiased estimator for the expected value of the maximum of a sample (largest order statistic) in the context of reporting experimental results, publishing a first-author paper at ACL 2020 [14].

Research Intern and Contractor May 2018–December 2019
Comcast Applied AI *Washington, District of Columbia*

• **Speech commands recognition**: implemented and productionized my undergraduate keyword spotting research for the X1 entertainment system, building a system that served more than 20 million voice queries per day [19, 39].

• **Speech recognition analysis**: advanced the understanding of speech recognition errors and their effects on the behavior of actual customers, publishing a first-author paper at SIGIR 2019 [18].

Undergraduate Research Assistant September 2017–December 2017
University of Waterloo *Waterloo, Ontario*

• **Small-footprint keyword spotting**: developed resource-efficient neural networks for keyword spotting, publishing two first-author papers at ICASSP 2018 [22, 23].

Awards and Scholarships

Best Paper Award, the 61st Annual Meeting of the Association for Computational Linguistics July 2023

Everyday Innovator Award, Comcast December 2022

Ideator Award, Comcast December 2023

Vector Institute Research Grant ($6,000) March 2020

GO-Bell Scholarship ($24,000) January 2019

Academic Mentorship

Jaejun (Brandon) Lee, Master's Thesis Student; Undergraduate 2018–2020
• Mentored research projects in keyword spotting, with papers at EMNLP 2019 [17] and IUI 2019 [20].

Ashutosh Adhikari, Master's Thesis Student 2018–2019
• Mentored research projects in document classification; papers at NAACL 2019 [21] and an ACL workshop [27].

Xinyu (Mavis) Liu, Undergraduate Research Assistant 2018
• Mentored an undergraduate research project in keyword spotting.

Education

Bachelor of Computer Science, Honours, University of Waterloo September 2013–December 2017

Program Committees AAAI (2020, 2021), AACL-IJCNLP (2020), ACL (2020, 2022, 2023), EACL (2021) EMNLP Industry Track (2023), NAACL (2021), ODML-CDNNR@ICML (2019)

Publications

Refereed Conference Proceedings

[1] Wenyan Li, Jiaang Li, Rita Ramos, **Raphael Tang**, Desmond Elliott. 2024. Understanding Retrieval Robustness for Retrieval-Augmented Image Captioning. *In ACL.*

[2] Scott Rome, Tianwen Chen, **Raphael Tang**, Luwei Zhou, Ferhan Ture. 2024. "Ask Me Anything": How Comcast Uses LLMs to Assist Agents in Real Time. *In SIGIR Industry.*

[3] **Raphael Tang***, Xinyu Zhang*, Xueguang Ma, Jimmy Lin, Ferhan Ture. 2024. Found in the Middle: Permutation Self-Consistency Improves Listwise Ranking in Large Language Models. *In NAACL.* (*equal contribution)

[4] Yao Lu, Jiayi Wang, **Raphael Tang**, Sebastian Riedel, Pontus Stenetorp. 2024. Strings from the Library of Babel: Random Sampling as a Strong Baseline for Prompt Optimisation. *In NAACL.*

[5] **Raphael Tang**,* Linqing Liu,* Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, Ferhan Ture. 2023. What the DAAM: Interpreting Stable Diffusion Using Cross Attention. *In ACL.* Winner of the *Best Paper Award*. (*equal contribution)

[6] Ji Xin, **Raphael Tang**, Zhiying Jiang, Yaoliang Yu, Jimmy Lin. 2023. Operator Selection and Ordering in a Pipeline Approach to Efficiency Optimizations for Transformers. *In Findings of ACL.*

[7] Zhiying Jiang, Matthew Y. R. Yang, Mikhail Tsirlin, **Raphael Tang**, Yiqin Dai, Jimmy Lin. 2023. "Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors. *In Findings of ACL.*

[8] **Raphael Tang**, Karun Kumar, Gefei Yang, Akshat Pandey, Yajie Mao, Vladislav Belyaev, Madhuri Emmadi, Craig Murray, Ferhan Ture, Jimmy Lin. 2022. SpeechNet: Weakly Supervised, End-to-End Speech Recognition at Industrial Scale. *In EMNLP, Industry Track.*

[9] **Raphael Tang**, Karun Kumar, Ji Xin, Piyush Vyas, Wenyan Li, Gefei Yang, Yajie Mao, Craig Murray, and Jimmy Lin. 2022. Temporal Early Exiting for Streaming Speech Commands Recognition. *In ICASSP.*

[10] **Raphael Tang**, Karun Kumar, Kendra Chalkley, Ji Xin, Liming Zhang, Wenyan Li, Gefei Yang, Yajie Mao, Junho Shin, Craig Murray, Jimmy Lin. 2021. Voice Query Auto Completion. *In EMNLP.*

[11] Ji Xin, **Raphael Tang**, Yaoliang Yu, Jimmy Lin. 2021. The Art of Abstention: Selective Prediction and Error Regularization for Natural Language Processing. *In ACL-IJCNLP.*

[12] Ji Xin, **Raphael Tang**, Yaoliang Yu, Jimmy Lin. 2021. BERxiT: Early Exiting for *BERT with Better Fine-Tuning and Extension to Regression. *In EACL.*

[13] Zhiying Jiang, **Raphael Tang**, Ji Xin, Jimmy Lin. 2020. Inserting Information Bottlenecks for Attribution in Transformers. *In Findings of EMNLP.*

[14] **Raphael Tang**, Jaejun Lee, Ji Xin, Xinyu Liu, Yaoliang Yu, Jimmy Lin. 2020. Showing Your Work Doesn't Always Work. *In ACL.*

[15] Ji Xin, **Raphael Tang**, Jaejun Lee, Yaoliang Yu, Jimmy Lin. 2020. DeeBERT: Dynamic Early Exiting for Accelerating BERT Inference *In ACL.*

[16] Linqing Liu, Wei Yang, Jinfeng Rao, **Raphael Tang**, Jimmy Lin. 2019. Incorporating Contextual and Syntactic Structures Improves Semantic Similarity Modeling. *In EMNLP-IJCNLP.*

[17] Jaejun Lee, **Raphael Tang**, Jimmy Lin. 2019. Honkling: In-Browser Personalization for Ubiquitous Keyword Spotting. *In EMNLP-IJCNLP 2019, System Demonstrations.*

[18] **Raphael Tang**, Ferhan Ture, Jimmy Lin. 2019. Yelling at Your TV: An Analysis of Speech Recognition Errors and Subsequent User Behavior on Entertainment Systems. *In SIGIR.*

[19] Ferhan Ture, Jinfeng Rao, **Raphael Tang**, Jimmy Lin. 2019. Challenges and Opportunities in Understanding Spoken Queries Directed at Modern Entertainment Platforms. *In SIGIR, Industry Track.*

[20] Jaejun Lee, **Raphael Tang**, Jimmy Lin. 2019. Universal Voice-Enabled User Interfaces using JavaScript. *In IUI, System Demonstrations.*

[21] Ashutosh Adhikari,* Achyudh Ram,* **Raphael Tang**, Jimmy Lin. 2019. Rethinking Complex Neural Network

Architectures for Document Classification. *In NAACL-HLT.* (*equal contribution)

[22] **Raphael Tang**, Weijie Wang, Zhucheng Tu, Jimmy Lin. 2018. An Experimental Analysis of the Power Consumption of Convolutional Neural Networks for Keyword Spotting. *In ICASSP.*

[23] **Raphael Tang**, Jimmy Lin. 2018. Deep Residual Learning for Small-Footprint Keyword Spotting. *In ICASSP.*

Refereed Workshop Proceedings

[24] Zhiying Jiang, **Raphael Tang**, Ji Xin, Jimmy Lin. 2021. How Does BERT Rerank Passages? An Attribution Analysis with Information Bottlenecks. *In BlackboxNLP at EMNLP.*

[25] **Raphael Tang**,* Jaejun Lee,* Afsaneh Razi, Julia Cambre, Ian Bicking, Jofish Kaye, Jimmy Lin. 2020. Howl: A Deployed, Open-Source Wake Word Detection System. *In NLPOSS at EMNLP.* (*equal contribution)

[26] Edwin Zhang, Nikhil Gupta, **Raphael Tang**, Xiao Han, Ronak Pradeep, Kuang Lu, Yue Zhang, Rodrigo Nogueira, Kyunghyun Cho, Hui Fang, Jimmy Lin. 2020. Covidex: Neural Ranking Models and Keyword Search Infrastructure for the COVID-19 Open Research Dataset. *In SDP at EMNLP.*

[27] Ashutosh Adhikari, Achyudh Ram, **Raphael Tang**, Jimmy Lin. 2020. Exploring the Limits of Simple Learners in Knowledge Distillation for Document Classification. *In Repl4NLP at ACL.*

[28] **Raphael Tang**, Yao Lu, Jimmy Lin. 2019. Natural Language Generation for Effective Knowledge Distillation. *In DeepLo at EMNLP-IJCNLP.*

[29] **Raphael Tang**, Ashutosh Adhikari, Jimmy Lin. 2018. FLOPs as a Direct Optimization Objective for Learning Sparse Neural Networks. *In CDNNRIA at NeurIPS.*

Patents

[30] **Raphael Tang**, Karun Kumar, Wenyan Li, Gefei Yang, Yajie Mao, Yang Hu, Rui Min, Geoffrey Murray. 2023. Methods and Systems for Voice Control. US Patent App. 17/566,036.

[31] **Raphael Tang**, Karun Kumar, Kendra Chalkley, Wenyan Li, Liming Zhang, Gefei Yang, Yajie Mao, Jun Ho Shin, Geoffrey Murray. 2023. Predictive Query Execution. US Patent App. 17/454,234.

[32] **Raphael Tang**, Karun Kumar, Kendra Chalkley, Wenyan Li, Liming Zhang, Gefei Yang, Yajie Mao, Liming Zhang, Jun Ho Shin, Pamela Shapiro, Geoffrey Murray. To be announced. Systems and Methods for Improved Automatic Speech Recognition Systems. US Patent App. 18/066174.

[33] **Raphael Tang**, Karun Kumar, Geoffrey Murray, Ferhan Ture. To be announced. BarraCUDA: Distribution-Aware CUDA Graph Pools for Performant Neural Inference. US Patent App. 18/193212.

Manuscripts

[34] **Raphael Tang**, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, Jimmy Lin. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19. arXiv.

[35] Jaejun Lee, **Raphael Tang**, Jimmy Lin. 2019. What Would Elsa Do? Freezing Layers During Transformer Fine-Tuning. arXiv.

[36] Yinan Zhang, **Raphael Tang**, Jimmy Lin. 2019. Explicit Pairwise Word Interaction Modeling Improves Pretrained Transformers for English Semantic Similarity Tasks. arXiv.

[37] Ashutosh Adhikari, Achyudh Ram, **Raphael Tang**, Jimmy Lin. 2019. DocBERT: BERT for Document Classification. arXiv.

[38] **Raphael Tang**,* Yao Lu,* Linqing Liu,* Lili Mou, Olga Vechtomova, Jimmy Lin. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. arXiv. (*equal contribution)

[39] **Raphael Tang**, Gefei Yang, Hong Wei, Yajie Mao, Ferhan Ture, Jimmy Lin. 2018. Streaming Voice Query Recognition using Causal Convolutional Recurrent Neural Networks. arXiv.

[40] **Raphael Tang**, Jimmy Lin. 2018. Progress and Tradeoffs in Neural Language Models. arXiv.

[41] Jaejun Lee, **Raphael Tang**, Jimmy Lin. 2018. JavaScript Convolutional Neural Networks for Keyword Spotting in the Browser: An Experimental Analysis. arXiv.

[42] **Raphael Tang**, Jimmy Lin. 2018. Adaptive Pruning of Neural Language Models for Mobile Devices. arXiv.

[43] **Raphael Tang**, Jimmy Lin. 2017. Honk: A PyTorch Reimplementation of Convolutional Neural Networks for Keyword Spotting. arXiv.

Selected Talks

What the DAAM: Interpreting Stable Diffusion Using Cross Attention. 2023. Presented at the best paper award ceremony of ACL 2023.

Comcast SpeechNet: Weakly Supervised, End-to-End Speech Recognition at Industrial Scale. 2023. Presented at Snorkel.AI's Future of Data-Centric AI.