Yuntian DENG

Google Scholar◇ Personal Website
yuntian@uwaterloo.ca

Employments

University of Waterloo — Waterloo, Canada
Assistant Professor, Cheriton School of Computer Science — Aug 2024 - Present

NVIDIA — Remote
Visiting Professor, Prof. Yejin Choi's team — Nov 2024 - Present

Harvard University — Remote
Associate, Computer Science — Jun 2024 - Present

Vector Institute — Toronto, Canada
Faculty Affiliate — Aug 2024 - Present

Allen Institute for Artificial Intelligence — Seattle, WA
Postdoc (advised by Prof. Yejin Choi) — Jul 2023 - Jul 2024

Education

Harvard University — Cambridge, MA
Ph.D. in Computer Science — May 2023
Advisors: Professors Alexander Rush and Stuart Shieber

Carnegie Mellon University — Pittsburgh, PA
Master in Language Technologies — Aug 2016
Advisor: Prof. Eric Xing

Tsinghua University — Beijing, China
Bachelor of Engineering, Department of Automation — Jul 2014

Awards

Argonne National Lab Impact Award — Jan 2023

University of Chicago Rising Stars in Data Science — Nov 2022

ACM Gordon Bell Prize — Nov 2022

NVIDIA Fellowship — Dec 2021

Microsoft Turing Academic Program, Selected for Improving LM Reasoning — Dec 2021

Harvard Certificates of Distinction in Teaching — Spring 2019, Fall 2020, Fall 2021

Twitch Fellowship Finalist	Jan 2021
DAC 2020 Best Paper Award	Jul 2020
Baidu Fellowship	Feb 2019
French American Doctoral Exchange Program (French Embassy, 10 US laureates)	Jul 2018
ACL 2017 Best Demo Paper Award Runner-Up	Aug 2017

Publications and Preprints

*Equal Contribution.

Selected Papers

34 **Yuntian Deng**, Yejin Choi, Stuart Shieber. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step. *arXiv 2024*

33 **Yuntian Deng**, Wenting Zhao, Jack Hessel, Xiang Ren, Claire Cardie, Yejin Choi. WildVis: Open Source Visualizer for Million-Scale Chat Logs in the Wild. *EMNLP 2024 Demo*

32 Bill Yuchen Lin, **Yuntian Deng**, Khyathi Chandu, Faeze Brahman, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, Yejin Choi. WildBench: Benchmarking LLMs with Challenging Tasks from Real Users in the Wild. *ICLR 2025* ***Spotlight***

31 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, **Yuntian Deng**. WildChat: 1M ChatGPT Interaction Logs In The Wild. *ICLR 2024* ***Spotlight, Covered by Washington Post and NZZ***

30 **Yuntian Deng**, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, Stuart Shieber. Implicit Chain of Thought Reasoning via Knowledge Distillation. *arXiv 2023*

29 John X. Morris, Chandan Singh, Alexander Rush, Jianfeng Gao, **Yuntian Deng**. Tree Prompting: Efficient Task Adaptation without Fine-Tuning. *EMNLP 2023*

28 **Yuntian Deng**, Volodymyr Kuleshov, Alexander Rush. Model Criticism for Long-Form Text Generation. *EMNLP 2022*

27 **Yuntian Deng**, Anton Bakhtin, Myle Ott, Arthur Szlam, Marc'Aurelio Ranzato. Residual Energy-Based Models for Text Generation. *ICLR 2020*

26 **Yuntian Deng**, Alexander Rush. Cascaded Text Generation with Markov Transformers. *NeurIPS 2020*

25 Sebastian Gehrmann, **Yuntian Deng**, Alexander Rush. Bottom-Up Abstractive Summarization. *EMNLP 2018*

24 Zachary Ziegler*, **Yuntian Deng***, Alexander Rush. Neural Linguistic Steganography. *EMNLP 2019* ***Oral*** (**Demo**: steganography.live)

23 **Yuntian Deng**, Noriyuki Kojima, Alexander Rush. Markup-to-Image Diffusion Models with Scheduled Sampling. *ICLR 2023* (Demo: huggingface.co/spaces/yuntian-deng/latex2im)

22 **Yuntian Deng***, Yoon Kim*, Justin Chiu, Demi Guo, Alexander Rush. Latent Alignment and Variational Attention. *NeurIPS 2018*

21 **Yuntian Deng**, Anssi Kanervisto, Jeffrey Ling, Alexander Rush. Image-to-Markup Generation with Coarse-to-Fine Attention. *ICML 2017* (Demo: im2markup.yuntiandeng.com)

Journal Papers

20 Anton Bakhtin*, **Yuntian Deng***, Sam Gross, Myle Ott, Marc'Aurelio Ranzato, Arthur Szlam. Residual Energy-Based Models for Text. *JMLR 2021*

Other Papers and Preprints

19 Yifan Zong, **Yuntian Deng**, Pengyu Nie. Mix-of-Language-Experts Architecture for Multilingual Programming. *LLM4Code 2025*

18 Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, **Yuntian Deng**, Yejin Choi. WildHallucinations: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries. *arXiv 2024*

17 Zhangchen Xu, Fengqing Jiang, Luyao Niu, **Yuntian Deng**, Radha Poovendran, Yejin Choi, Bill Yuchen Lin. Magpie: Alignment Data Synthesis from Scratch by Prompting Aligned LLMs with Nothing. *ICLR 2025*

16 Jinjie Ni, Fuzhao Xue, Xiang Yue, **Yuntian Deng**, Mahir Shah, Kabir Jain, Graham Neubig, Yang You. MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. *NeurIPS 2024*

15 Richa Rastogi, Yair Schiff, Alon Hacohen, Zhaozhi Li, Ian Lee, **Yuntian Deng**, Mert R. Sabuncu, Volodymyr Kuleshov. Semi-Parametric Inducing Point Networks and Neural Processes. *ICLR 2023*

14 Maxim Zvyagin*, Alexander Brace*, Kyle Hippe*, **Yuntian Deng***, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan, GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. **ACM Gordon Bell Special Covid Prize**

13 Justin Chiu*, **Yuntian Deng***, Alexander Rush. Low-Rank Constraints for Fast Inference in Structured Models. *NeurIPS 2021*

12 **Yuntian Deng**, Alexander Rush. Sequence-to-Lattice Models for Fast Translation. *EMNLP 2021 Findings*

11 Keyon Vafa, **Yuntian Deng**, David Blei, Alexander Rush. Rationales for Sequential Predictions. *EMNLP 2021* ***Oral***

10 Anton Bakhtin, Sam Gross, Myle Ott, **Yuntian Deng**, Marc'Aurelio Ranzato, Arthur Szlam. Real or Fake? Learning to Discriminate Machine from Human Generated Text. *arXiv 1906.03351*

9 Thierry Tambe, En-Yu Yang, Zishen Wan, **Yuntian Deng**, Vijay Janapa Reddi, Alexander Rush, David Brooks, Gu-Yeon Wei. Algorithm-Hardware Co-Design of Adaptive Floating-Point Encodings for Resilient Deep Learning Inference. *Design Automation Conference (DAC) 2020* ***Best Paper Award***

8 **Yuntian Deng**, David Rosenberg, Gideon Mann. Challenges in End-to-End Neural Scientific Table Recognition. *ICDAR 2019*

7 Pengtao Xie, **Yuntian Deng**, Yi Zhou, Abhimanu Kumar, Yaoliang Yu, James Zou, Eric P Xing. Learning Latent Space Models with Angular Constraints. *ICML 2017*

6 Guillaume Klein, Yoon Kim, **Yuntian Deng**, Jean Senellart, Alexander M Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ACL 2017* ***Best Demo Paper Award Runner-Up***

5 Zichao Yang, Zhiting Hu, **Yuntian Deng**, Chris Dyer, Alex Smola. Neural Machine Translation with Recurrent Attention Modeling. *EACL 2017*

4 Xuezhe Ma, Yingkai Gao, Zhiting Hu, Yaoliang Yu, **Yuntian Deng**, Eduard Hovy. Dropout with Expectation-linear Regularization. *ICLR 2017*

3 Hao Zhang, Zhiting Hu, **Yuntian Deng**, Mrinmaya Sachan, Zhicheng Yan, Eric P. Xing. Learning Concept Taxonomies from Multi-modal Data. *ACL 2016*

2 Pengtao Xie, **Yuntian Deng**, Eric P. Xing. Diversifying Restricted Boltzmann Machine for Document Modeling. *KDD 2015*

1 Zhiting Hu, Poyao Huang, **Yuntian Deng**, Yingkai Gao, Eric P. Xing. Entity Hierarchy Embedding. *ACL 2015*

Media Coverage

NZZ: What people are really asking Chat-GPT?	Nov 30, 2024
TechCrunch: Why is ChatGPT so bad at math?	Oct 2, 2024
The Washington Post: What do people really ask chatbots?	Aug 4, 2024

Teaching Experience

University of Waterloo

Instructor, Intro to AI	May 2025 - Dec 2025
Guest Lecture, Computational Linguistics (Prof. Freda Shi)	Feb 11, 2025
Instructor, Topics in Language Models	Aug 2024 - Dec 2024

Harvard University

Head TA, Intro to Computational Linguistics and NLP (Prof. Stuart Shieber)	Aug 2021 - Dec 2021
Head TA, Intro to Computational Linguistics and NLP (Prof. Stuart Shieber)	Aug 2020 - Dec 2020
TA, Natural Language Processing (Alexander Rush)	Jan 2019 - May 2019
Lecture, Natural Language Processing - Translation (Prof. Alexander Rush)	Feb 13, 2019
TA, Advanced Machine Learning (Alexander Rush)	Aug 2017 - Dec 2017

Cornell University

Lab Material Preparation, Break Through Tech AI (Prof. Alexander Rush)	Aug 2021

Carnegie Mellon University

TA, Probabilistic Graphical Models (Prof. Matthew Gormley, Eric Xing)	Jan 2016 - May 2016
TA, Graduate Machine Learning (Profs. Ziv Bar Joseph, Eric Xing)	Aug 2015 - Dec 2015

Professional Service

2025: Area Chair of ICLR, ACL ARR, COLM, Reviewer of TACL, ICML, NeurIPS

2024: Area Chair of ICLR, ACL ARR, Oral Session Chair of ICLR, EMNLP, Reviewer of TACL, ICML, ACL ARR, COLM, COLING, ACL Demo

2023: Area Chair of EMNLP GEM Workshop, Reviewer of TACL, ICML, ACL, ACL ARR, ACL Demo, AACL, EMNLP, Foundations and Trends in Signal Processing, Computational Linguistics

2022: Reviewer of ACL ARR, EMNLP, ICML, NeurIPS, ICLR, EACL, ML Reproducibility Challenge, ACL Demo, NAACL Demo, and EMNLP GEM Workshop.

2021: Expert Reviewer of ICML. Oral Session Volunteer of ACL. Reviewer of NeurIPS, ACL, ACL ARR, ICLR, AAAI, EACL, NAACL, ML Reproducibility Challenge, ICLR EBM Workshop, IEEE TIIS, and NEJLT. Volunteer of EMNLP.

2020: Top 10% Reviewer of NeurIPS. Reviewer of ICML, ACL, EMNLP, AACL, AAAI, ACL Demo, COLING, IEEE TIFS, Transactions on Information Systems, and Journal of Computer Science and Technology. Volunteer of ICML. Volunteer moderator of ACL.

2019: Reviewer of NeurIPS, EMNLP, ICLR, NAACL, AAAI, IEEE TNNLS, ACM Computing Surveys, NeurIPS LIRE Workshop, NeurIPS Reproducibility Challenge, EMNLP Summarization Workshop, ICML GraphReason Workshop, and ICLR DeepGenStruct Workshop.

Community Service

Co-Treasurer, Board of Directors, Bright Starts Early Learning Centre	May 2024 - May 2026

Childcare Chair at NAACL 2022 D&I Committee	Apr - Jul 2022
Leader of Harvard English Language Table	Sep 2021 - Jan 2023
Mentor of Harvard Women in CS NLP Reading Group	Sep 2020 - Sep 2023
Volunteer Instructor of AP CS at Revere High School	Sep 2020 - May 2021

Internships

NVIDIA (Dr. Anima Anandkumar/Dr. Weili Nie/Dr. Arash Vahdat/Dr. Chaowei Xiao)	May - Dec '22
Facebook AI Research (Dr. Marc'Aurelio Ranzato, Dr. Arthur Szlam)	May - Dec 2019
Bloomberg CTO Office (Dr. David Rosenberg, Dr. Gideon Mann)	Jan - Aug 2017
UCSD (Dr. Charles Elkan)	Jul - Sep 2013

Talks

Google DeepMind: Implicit Reasoning in Language Models	May 2025, Remote
Colin Raffel Group: Implicit Reasoning in Language Models	Mar 2025, Toronto
U Oklahoma: Implicit Chain of Thought Reasoning	Jan 2025, Remote
NYU: Implicit Chain of Thought Reasoning	Nov 2024, New York, NY
Google DeepMind: Implicit Chain of Thought Reasoning	Jul 2024, Remote
Queen's University: Implicit Chain of Thought Reasoning	Feb 2024, Remote
UBC: Implicit Chain of Thought Reasoning	Jan 2024, Vancouver, Canada
AI2 Mosaic: WildChat	Sep 2023, Seattle, Washington
AI2 Semantic Scholars: Structure Modeling in Language Models	Aug 2023, Seattle, Washington
NYU AI School 2023: Natural Language Processing	Jun 2023, New York, NY
AI2 Mosaic: Structure Modeling in Language Models	Jun 2023, Seattle, Washington
UMass Amherst: Structural Coherence in Text Generation	May 2023, remote
Georgia Tech: Structural Coherence in Text Generation	Apr 2023, Atlanta, Georgia
Harvard Kempner Institute: Structural Coherence in Text Generation	Mar 2023, remote
U Alberta: Structural Coherence in Text Generation	Mar 2023, Edmonton, Canada
TTIC: Structural Coherence in Text Generation	Feb 2023, Chicago
U Waterloo ECE: Structural Coherence in Text Generation	Feb 2023, remote

Cornell Seminar in NLU: Structural Coherence in Text Generation	Feb 2023, New York
U Waterloo CS: Structural Coherence in Text Generation	Jan 2023, Waterloo, Canada
U Chicago Rising Star: Model Criticism for Long-Form Text Generation	Nov 2022, Chicago
Princeton NLP: Model Criticism for Long-Form Text Generation	Nov 2022, Princeton
EMNLP 2021 Findings: Sequence-to-Lattice Models for Fast Translation	Oct 2021, remote
OpenAI: Residual Energy-based Models for Text Generation	Apr 2021, remote
NeurIPS 2020: Cascaded Text Generation with Markov Transformers	Oct 2020, remote
Baidu: Cascaded Text Generation with Markov Transformers	Jun 2020, remote
ICLR 2020: Residual Energy-based Models for Text Generation	Apr 2020, remote
LinkedIn: Residual Energy-based Models for Text Generation	Mar 2020, remote
Wayfair: Neural Encoder-Decoder Models	Nov 2019, Boston
EMNLP 2019: Neural Linguistic Steganography	Nov 2019, Hong Kong
FAIR NLP Meeting: Energy-Based Models for Text Generation	Aug 2019, New York City
NeurIPS 2018: Latent Alignment and Variational Attention	Dec 2018, Montreal, Canada
The French National Center for Scientific Research: Variational Attention	Jul 2018, France
Association for Machine Translation in the Americas: OpenNMT	Mar 2018, Boston
Open-Source Neural Machine Translation Workshop: Image-to-Text	Mar 2018, Paris, France
Tencent Social Network Group TSAIC: OpenNMT	Dec 2017, Shenzhen, China
ICML: Image-to-Markup Generation	Aug 2017, Sydney, Australia

Panels

Vector Institute's Remarkable 2025 Mar 2025, Toronto
Panelist alongside Colin Raffel, Frank Rudzicz, Gerald Penn, Hongyang Zhang, Tracy Jenkin, and David Duvenaud. Discussed questions in NLP.

ACL Mentorship: Sharing Learnings & Identifying Research Directions Dec '23, Singapore
Panelist alongside Mohit Bansal, Zhijing Jin, and Rada Mihalcea. Discussed promising research directions.

ACL Mentorship: Preparing Your Grad School Application Oct 2023, Remote
Participated in a mentorship session with Luke Zettlemoyer, Zhijing Jin, Nathan Schneider, and Oana Ignat, focused on preparing grad school applications.

NAACL 2022: Balancing Professional and Family Commitments Jul 2022, Remote
Co-hosted with Shirley Hayati, discussing the balance between professional and family commitments.

Thesis Committees

- Member, MMath Thesis Defense Committee for Achint Soni (University of Waterloo, advised by Prof. Sirisha Rambhatla and Charles Clarke) May 2025
- Member, MMath Thesis Defense Committee for Chi-Chung Cheung (University of Waterloo, advised by Prof. Anita Layton) Dec 2024
- Member, PhD Thesis Proposal Committee for William Loh (University of Waterloo, advised by Prof. Pascal Poupart) Aug 2024

Open Source Projects

Internalize CoT Step by Step (171 Github stars)	PyTorch
WildChat o1 Chatbot (453 Hugging Face stars)	PyTorch
WildChat o1-mini Chatbot (261 Hugging Face stars)	PyTorch
WildChat GPT-4 Chatbot (493 Hugging Face stars)	PyTorch
WildChat GPT-3.5 Chatbot (196 Hugging Face stars)	PyTorch
Cascaded Generation (129 Github stars)	PyTorch
Neural Lingustic Steganography (182 Github stars)	PyTorch
Variational Attention (325 Github stars)	PyTorch
Image-to-Markup Generation (1.2k Github stars)	LuaTorch
OpenNMT-py (6.7k Github stars)	PyTorch
OpenNMT (2.4k Github stars)	LuaTorch
Attention OCR (1.1k Github stars)	Tensorflow