TRISTAN THRUSH

http://www.tristanthrush.com
https://github.com/TristanThrush
https://scholar.google.com/citations?user=qDDmq54AAAAJ

EDUCATION

Stanford University *2023 - Present*
PhD in Computer Science, Natural Language Processing Group
Rotation Supervisors: Christopher Potts, Tatsunori Hashimoto, Diyi Yang

Massachusetts Institute of Technology *2017 - 2019*
Master of Engineering in Computer Science with concentration in Artificial Intelligence
Thesis: SAL: a Self-Aware Learning system
Supervisor: Patrick Winston, Ford Professor of Artificial Intelligence and Computer Science
GPA: 5.0 (5.0 scale)

Massachusetts Institute of Technology *2015 - 2019*
Bachelor of Science, Computer Science and Engineering. Minor in Mathematics. Minor in Linguistics.
GPA: 4.7 (5.0 scale)

Graduated high school a year early as valedictorian after taking 7 engineering and upper division math classes at UCSD; attended MIT a year early as a result.

RESEARCH EXPERIENCE

Stanford AI Lab

- PhD Student *2023 - Present*
  - Supervising and leading a variety of projects around scaling laws, multimodality, datasets, and evaluation [2] [1]

Contextual AI

- Founding Member of Technical Staff *Summer 2023*
  - Lead for retrieval augmented LLM eval pipeline before joining Stanford in the fall
  - Research on Multimodal LLMs [3]

Hugging Face

- Research Engineer *2022 - 2023*
  - Contributor to the Hugging Face Reinforcement Learning from Human Feedback (RLHF) repo (see: `https://github.com/lvwerra/trl` and `https://github.com/lvwerra/trl`) and lead for crowdworker RLHF data collection.
  - Benchmarking of distributed frameworks for multi-node large language model training, to compare to our in-house solutions.
  - Lead for the Online Language Modelling (OLM) project [5], which is about continuous training of large language models for every new Common Crawl snapshot. Developed highly parallelized data generation tools that pull Common Crawl data, and filter for language, quality, content, recency, and duplicates. (See: `https://github.com/huggingface/olm-datasets` and `https://github.com/huggingface/olm-training`.)
  - Contributed to the pretraining data pipeline and the model card for the 176 billion parameter BigScience/BLOOM model [9] [7]
  - Added statistical bootstrap code and metrics to the Evaluate repo (see: `https://github.com/huggingface/evaluate`).
  - Implemented pipelines for the evaluation of huge language models on various dataset types for the Evaluation on the Hub project [8]. Involved the development of the zero shot evaluation

feature used for Anthropic's inverse scaling competition (see: `https://huggingface.co/blog/zero-shot-eval-on-the-hub`).
    - Added dataset perplexity measurements and other features for the Data Measurements Tool (see: `https://huggingface.co/spaces/huggingface/data-measurements-tool`)
    - Implemented a connection between Hugging Face Spaces and Amazon Mechanical Turk for easy crowdworker data collection

ML Commons

- Contributor *2022*
    - Supervision of engineers continuing the Dynabench project (see below), after it moved from Facebook AI to ML Commons
    - Engineering for DataPerf core-set selection research [10]

Facebook AI Research

- Research Associate *2020 - 2022*
    - Supervisor: Douwe Kiela, then Adina Williams
    - Engineering lead for the Dynabench AI benchmarking project [21] [19] [15]. Several governments (UK, EU) and companies (Facebook, Google, Microsoft, Huawei, Tencent, Wikipedia, Amazon, ML Commons, etc.) impacted, 3000+ users, 50+ model evaluation datasets hosted, 5+ large datasets generated, 10+ research papers enabled, 500+ models uploaded, 1 large scale machine translation competition hosted. (See: `https://dynabench.org/`.)
    - Lead for the Winoground vision+language model evaluation dataset [14]. At the time of this writing, it is the most downloaded Facebook dataset on Hugging Face (see: `https://huggingface.co/datasets/facebook/winoground`), beating FLORES, MultiLingual LibriSpeech, and PMD. It is also in the top 2% of all Hugging Face datasets by downloads and top 0.2% by likes.
    - Research on hate speech detection [23] [11], question answering [20] [12], natural language inference [16], and visual question answering [18]
    - Engineering for the WMT 2021 machine translation competition [17], and Facebook teams working on hate speech detection in production.

MIT Brain and Cognitive Sciences

- Research Associate, Computational Psycholinguistics Lab *2019 - 2020*
    - Supervisor: Roger Levy
    - Led research on few-shot learning in foundation language models [24], and cognitively inspired neural machine translation models [25]

MIT Computer Science and Artificial Intelligence Lab

- Undergraduate Researcher then Graduate Researcher, Genesis Group *2016 - 2019*
    - Supervisor: Patrick Winston, Randall Davis
    - Cognitive AI and reinforcement learning research [29] [26]
- Undergraduate Researcher, Robot Locomotion Group *Summer 2017*
    - Supervisor: Russ Tedrake
    - Integrated Leslie Kaelbling's BHPN planner with the Drake robot simulation and control toolbox. Then, integrated BHPN with a language production system to explain itself.
    - Contributed to Drake tools (see: `https://github.com/RobotLocomotion/drake`)

NASA/Caltech Jet Propulsion Lab

- Research Intern, Perception Systems Group *Summer 2018*
    - Supervisor: Renaud Detry
    - Developed a stereo visual odometry algorithm for sample tube localization on Mars. Involved research on neural dense segmentation and feature extraction modules. Developed an asso-

ciated dataset and data collection tools. Impacted the group's approach to autonomously bringing soil samples from Mars to Earth for the first time in human history. [22]
    - After the internship, remained as a remote affiliate until the end of 2018 to help a postdoc take over the work.
- Research Extern, Perception Systems Group *Winter 2017*
  - Supervisor: Renaud Detry
  - Combined inverse kinematics, motion planning, and vision systems for a robot arm
  - System was used as part of a project that won the best paper award at IROS

PUBLICATIONS

Available PDFs are here

1. ColorSwap: A Color and Word Order Dataset for Multimodal Models.
   Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, Diyi Yang, **Tristan Thrush**.
   In progress, 2024.

2. I am a Strange Dataset: Metalinguistic Tests for Language Models.
   Tristan Thrush, Jared Moore, Miguel Monares, Christopher Potts, Douwe Kiela.
   arXiv, 2024.

3. Towards Language Models That Can See: Computer Vision Through the LENS of Natural Language.
   William Berrios, Gautam Mittal, **Tristan Thrush**, Douwe Kiela, Amanpreet Singh.
   arXiv, 2023.

4. TRL: Transformer Reinforcement Learning.
   Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, **Tristan Thrush**, Nathan Lambert, Shengyi Huang.
   GitHub Repo, 2023. `https://github.com/huggingface/trl`

5. Online Language Modelling Dataset Pipeline.
   Tristan Thrush, Helen Ngo, Nathan Lambert, and Douwe Kiela.
   Github Repo, 2023. `https://github.com/huggingface/olm-datasets`

6. Measuring Data.
   Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, **Tristan Thrush**, Yacine Jernite, Douwe Kiela.
   arXiv, 2022.

7. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.
   Teven Le Scao et al. (approx 400 authors).
   BigScience Workshop, 2022.

8. Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurement.
   Leandro von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, **Tristan Thrush**, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, Omar Sanseviero, Mario Šaško, Albert Villanova, Quentin Lhoest, Julien Chaumond, Margaret Mitchell, Alexander M Rush, Thomas Wolf, and Douwe Kiela.
   EMNLP System Demos, 2022.

9. The BigScience ROOTS Corpus: A 1.6 TB Composite Multilingual Dataset.
   Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor,

Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, **Tristan Thrush**, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Romero Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Vu Minh Chien, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Luccioni, and Yacine Jernite.
NeurIPS Datasets and Benchmarks, 2022.

10. DataPerf: Benchmarks for Data-Centric AI Development.
Mark Mazumder, Colby Banbury, Xiaozhe Yao, Bojan Karlaš, William Gaviria Rojas, Sudnya Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Sabri Eyuboglu, Amirata Ghorbani, Emmett Goodman, Tariq Kane, Christine R Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, **Tristan Thrush**, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Ng, Peter Mattson, and Vijay Janapa Reddi.
White Paper, 2022.

11. Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate.
Hannah Rose Kirk, Bertram Vidgen, Paul Röttger, **Tristan Thrush**, and Scott A Hale.
NAACL, 2022.

12. Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants.
Max Bartolo, **Tristan Thrush**, Sebastian Riedel, Pontus Stenetorp, Robin Jia, and Douwe Kiela.
NAACL, 2022.

13. Proceedings of the First Workshop on Dynamic Adversarial Data Collection.
Max Bartolo, Hannah Kirk, Pedro Rodriguez, Katerina Margatina, **Tristan Thrush**, Robin Jia, Pontus Stenetorp, Adina Williams, and Douwe Kiela.
DADC at NAACL, 2022.

14. Winoground: Probing vision and language models for visio-linguistic compositionality.
Tristan Thrush*, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross*.
CVPR, 2022.

15. Dynatask: A framework for creating dynamic AI benchmark tasks.
Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gaviria Rojas, Peter Mattson, Adina Williams, and Douwe Kiela.
ACL System Demos, 2022.

16. ANLIzing the Adversarial Natural Language Inference Dataset.
Adina Williams, **Tristan Thrush**, and Douwe Kiela.
SCiL, 2022.

17. Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation.
Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, **Tristan Thrush**, and Francisco Guzmán.
WMT at EMNLP, 2021.

18. Human-Adversarial Visual Question Answering.
Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Alberto Lopez Magana, **Tristan Thrush**, Wojciech Galuba, Devi Parikh, and Douwe Kiela.
NeurIPS, 2021.

19. Dynaboard: An Evaluation-As-A-Service Platform for Holistic Next-Generation Benchmarking.

Zhiyi Ma*, Kawin Ethayarajh*, **Tristan Thrush***, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela.
NeurIPS, 2021.

20. Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation.
Max Bartolo, **Tristan Thrush**, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela.
EMNLP, 2021.

21. Dynabench: Rethinking benchmarking in NLP.
Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, **Tristan Thrush**, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams.
NAACL, 2021.

22. Rover Relocalization for Mars Sample Return by Virtual Template Synthesis and Matching.
Tu-Hoa Pham, William Seto, Shreyansh Daftry, Barry Ridge, Johanna Hansen, **Tristan Thrush**, Mark Van der Merwe, Gerard Maggiolino, Alexander Brinkman, John Mayo, Yang Cheng, Curtis Padgett, Eric Kulczycki, and Renaud Detry.
IEEE Robotics and Automation Letters, 2021.

23. Learning from the worst: Dynamically generated datasets to improve online hate detection.
Bertie Vidgen, **Tristan Thrush**, Zeerak Waseem, Douwe Kiela.
ACL, 2021.

24. Investigating Novel Verb Learning in BERT: Selectional Preference Classes and Alternation-Based Syntactic Generalization.
Tristan Thrush, Ethan Wilcox, and Roger Levy.
BlackboxNLP at EMNLP, 2020.

25. Compositional neural machine translation by removing the lexicon from syntax (Abstract).
Tristan Thrush.
CogSci, 2020.

26. SAL: a Self-Aware Learning system (Master's thesis).
Tristan Thrush.
MIT Computer Science and Artificial Intelligence Laboratory, 2019.

27. Convolutions inspired by the human retina enable learning of more robust features.
Tristan Thrush.
MIT Department of Brain and Cognitive Sciences, 2018.

28. A neural model for learning a humanlike vowel feature space.
Tristan Thrush.
Northeastern Computational Phonology, 2018.

29. The partial mental state inducer: Learning intuition with few training examples and k-line theory.
Tristan Thrush, and Patrick Winston.
Advances in Cognitive Systems, 2018.

30. Machine learning approaches to capture the reliability of news articles.
Rares Buhai*, and **Tristan Thrush***.
MIT Department of Electrical Engineering and Computer Science, 2017.

31. A self-aware and hypothetical question-answering BHPN with drake control.
Tristan Thrush.
MIT Computer Science and Artificial Intelligence Laboratory, 2017.

32. Probabilistic lattice learning and backward chaining.
Tristan Thrush.
MIT Computer Science and Artificial Intelligence Laboratory, 2016.

HONORS

Stanford Graduate Fellowship (SGF)	*2023 - Present*
Winoground [14] named in the 100 most cited AI papers of 2022	*2023*
Evaluate & Evaluation on the Hub [8] won Best Demo Paper, EMNLP	*2022*
Funding from grants, MIT Computational Psycholinguistics Lab	*2019 - 2020*
MEng thesis proposal was selected for RA funding	*2018 - 2019*
Certificate in Advanced Undergrad Research in AI and Machine Learning, MIT	*2018*
Author of three research proposals accepted for MIT lab sponsored funding	*2017 - 2018*
MIT EECS Undergraduate Research and Innovation Scholar	*2017 - 2018*
Author of research proposal accepted for MIT institute funding	*2017*

TALKS AND PRESENTATIONS

ICWSM (Invited Talk)	*2023*
CVPR, New Orleans	*2022*
ACL, Dublin	*2022*
NeurIPS, Virtual	*2021*
BlackboxNLP, Virtual	*2020*
CogSci, Virtual	*2020*
Northeastern Computational Phonology, MIT	*2019*
Advances in Cognitive Systems, Stanford	*2018*

SERVICE

Workshops Organized

Dynamic Adversarial Data Collection (DADC) at NAACL	*2022*
Conference on Machine Translation (WMT) at EMNLP	*2021*

Area Chair

NAACL	*2024*

Reviewing

EACL	*2023*
NeurIPS	*2022, 2023*
Dynamic Adversarial Data Collection (DADC) at NAACL	*2022*
Dataperf at ICML	*2022*
Workshop on Online Abuse and Harms (WOAH) at ACL	*2021, 2022*
Data-Centric AI at NeurIPS	*2021*
Conference on Machine Translation (WMT) at EMNLP	*2021*

TEACHING

MIT Brain and Cognitive Sciences

- Led NLP tutorial for undergrads, Computational Psycholinguistics Group *Spring 2020*
- Teaching Assistant, Computational Cognitive Science Class *Fall 2019*
  - Supervisor: Josh Tenenbaum
  - Supervised grad students' NLP research projects, focusing on grammar induction

- Received highest possible ratings from student evaluations at the end of the course
- Other normal TA responsibilities: helped students at office hours, recorded lectures, etc.

MENTORING

Undergrads and Masters Students

- Franklin Wang, MIT *2023 - Present*
- Neil Chowdhury, MIT, now on leave for OpenAI *2023 - Present*
- Sumedh Shenoy, MIT *2023 - Present*
- Jirayu Burapacheep, Stanford *2023 - Present*
- Ishan Gaur, Stanford *2023 - Present*
- Agam Bhatia, Stanford *2023 - Present*
- Miguel Monares, UCSD, now cofounder at Playtest AI *2023*

Major League Hacking fellows working on Facebook AI projects

- Ishita Dasgupta *2021*
- Anand Rajaram *2021*
- Wong Kok Rui *2021*
- Fatima Zahra Chriha *2021*

Facebook AI SWE interns

- Anmol Gupta (partial supervision), now full SWE at Meta *2021*

SELECTED PRESS AND MEDIA

Pop Culture / Pop Science: Coexisting with AI podcast

Winoground [14]: Gary Marcus (#3 on Hacker News), Jack Clark/Import AI, Stanford Podcast, Tweet

Online Language Modelling [5]: Jack Clark/Import AI, Tweet

BigScience and BLOOM Language Model [9]: Washington Post

Dynabench [21]: Facebook AI Blog, MIT Tech Review, Wired, Market Tech Post, NLP Highlights

LANGUAGES AND FRAMEWORKS

Huggingface Transformers. Frameworks for very large distributed model training (e.g. DeepSpeed). Large-Scale Data Processing. Amazon Mechanical Turk experiment design. Mephisto. Python. Cython. C++. C. Java. JavaScript. Lua. R. MATLAB. React js. WebPPL. Assembly. SQL. YAML. Batch Files. Shell Scripts. XML-ish (HTML, URDF, SDF). AWS. GCP. Production Model Deployment. Cluster Computing and Machine Learning. ROS. Drake. LCM. PyTorch. Torch. Keras. TensorFlow. OpenCV. OpenAI Gym. Pandas. scikit-learn. NumPy. SciPy. GNU/UNIX. Android. LIBVISO2. Linguistic Research Databases (e.g. CHILDES, Treebanks, VerbNet).