Ethan Perez | CV

ethan@anthropic.com • ethanperez.net • ethanjperez

Current Positions

Anthropic — **Research Scientist**
AI Safety Team — *May 2022–*

Fund for Alignment Research — **Board Member**
2021-

Education

New York University — **Ph.D. Computer Science**
GPA: 4.0. Advisors: Kyunghyun Cho and Douwe Kiela — *September 2018–March 2022*

Rice University — **B.A. Computer Science**
GPA: 3.98. Distinction in Research and Creative Work — *August 2014–May 2018*

Papers

"Few-shot Adaptation Works with UnpredicTable Data" *arXiv 2022*.
Jun Shern Chan, Michael Pieler, Jonathan Jao, Jérémy Scheurer, **Ethan Perez**.

"Language Models (Mostly) Know What They Know" *arXiv 2022*.
Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, **Ethan Perez**, ..., Ben Mann, Sam McCandlish, Chris Olah, Jared Kaplan.

"RL with KL penalties is better viewed as Bayesian Inference."
RL as a Model of Agency workshop @ RLDM 2022.
Tomasz Korbak, **Ethan Perez**, Christopher L Buckley.

"Training Language Models with Natural Language Feedback."
ACL 2022 Workshop on Learning with Natural Language Supervision.
Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, **Ethan Perez**.

"Single-Turn Debate Does Not Help Humans Answer Hard Reading-Comprehension Questions."
ACL 2022 Workshop on Learning with Natural Language Supervision.
Alicia Parrish, Harsh Trivedi, **Ethan Perez**, Angelica Chen, Nikita Nangia, Jason Phang, Samuel R. Bowman.

"Finding and Fixing Undesirable Behaviours in Pretrained Language Models." *PhD Thesis 2022*.
Ethan Perez.

"Red Teaming Language Models with Language Models." *arXiv 2022*.
Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, Geoffrey Irving.

"True Few-shot Learning with Language Models." *NeurIPS 2021*.
Ethan Perez, Douwe Kiela, Kyunghyun Cho.

"Case-based Reasoning for Natural Language Queries over Knowledge Bases." *EMNLP 2021*.
Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, **Ethan Perez**, Jay-Yoon Lee, Lizhen Tan,

Lazaros Polymenakos, Andrew McCallum.

"Rissanen Data Analysis: Examining Dataset Characteristics via Description Length." *ICML 2021*.
Ethan Perez, Douwe Kiela, Kyunghyun Cho.

"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *NeurIPS 2020*.
Patrick Lewis, **Ethan Perez**, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela.

"Unsupervised Question Decomposition for Question Answering." *EMNLP 2020*.
Ethan Perez, Patrick Lewis, Wen-tau Yih, Kyunghyun Cho, Douwe Kiela.

"Supervised Multimodal Bitransformers for Classifying Images and Text." *arXiv 2020*.
Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, **Ethan Perez**, Davide Testuggine.

"Finding Generalizable Evidence by Learning to Convince Q&A Models." *EMNLP 2019*.
Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, Kyunghyun Cho.

"Retrospective for 'FiLM: Visual Reasoning with a General Conditioning Layer."'
NeurIPS ML Retrospectives Workshop 2019.
Ethan Perez.

"ELI5: Long Form Question Answering." *ACL 2019*.
Angela Fan, Yacine Jernite*, **Ethan Perez***, David Grangier, Michael Auli, Jason Weston.

"Visual Reasoning with Multi-hop Feature Modulation." *ECCV 2018*.
Florian Strub, Mathieu Seurin, **Ethan Perez**, Harm de Vries, Jérémie Mary,
Philippe Preux, Aaron Courville, Olivier Pietquin.

"Feature-wise Transformations." *Distill 2018*.
Vincent Dumoulin, **Ethan Perez**, Nathan Schucher, Florian Strub,
Harm de Vries, Aaron Courville, Yoshua Bengio.

"HoME: a Household Multimodal Environment."
NIPS 2017 Workshop on Visually-Grounded Interaction and Language Workshop.
Simon Brodeur, **Ethan Perez***, Ankesh Anand*, Florian Golemo*, Luca Celotti,
Florian Strub, Jean Rouat, Hugo Larochelle, Aaron Courville.

"FiLM: Visual Reasoning with a General Conditioning Layer." *AAAI 2018*.
Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, Aaron Courville.

"Learning Visual Reasoning Without Strong Priors."
ICML 2017 Workshop on Machine Learning for Speech and Language Processing.
Ethan Perez, Harm de Vries, Florian Strub, Vincent Dumoulin, Aaron Courville.

"Semi-Supervised Learning with the Deep Rendering Mixture Model." *arXiv 2016*.
Tan Nguyen, Wanjia Liu, **Ethan Perez**, Richard G. Baraniuk, Ankit B. Patel.

Research & Industry Experience

DeepMind: Aligning Language Models with Human Preferences
Research with Geoffrey Irving — *Summer 2021*

Facebook AI Research: Multi-hop Question Answering
Research with Douwe Kiela, Kyunghyun Cho, and Wen-tau Yih — *Summer 2019–Spring 2020*

Facebook AI Research: Long-form Question Answering
Research with Jason Weston, Michael Auli, and David Grangier — *Summer 2018*

University of Montreal: Visual Question Answering
Research with Aaron Courville and Hugo Larochelle *Summer 2017–Spring 2018*

Rice University: Semi-Supervised Image Classification
Research with Ankit Patel *Fall 2016–Spring 2017*

Uber: Fraud Detection with Machine Learning
Internship via KPCB Fellowship *Summer 2016*

Google Maps: Localization Algorithms
Internship *Summer 2015*

Invited Talks

"True Few-shot Learning with Language Models."
DeepMind Large-Scale Deep Learning Group, June 2021.

"True Few-shot Learning with Language Models."
Berkeley Center for Human Compatible AI Seminar, June 2021.

"Rissanen Data Analysis: Examining Dataset Characteristics with Description Length."
DeepMind and Future of Humanity Institute AI Safety Talk Series, July 2021.

"Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks."
DeepMind Natural Language Processing Group, November 2020.

"Proof of Concepts in Aligned Question-Answering."
DeepMind and Future of Humanity Institute AI Safety Talk Series, July 2020.

"FiLM: Visual Reasoning with a General Conditioning Layer."
Montreal AI Symposium, September 2017.
`https://youtu.be/02xIkHowQOk?t=2h44m55s`

"Learning Visual Reasoning Without Strong Priors."
Deep Language Workshop, University of Montreal, September 2017.

Awards

Open Philanthropy AI Alignment Research Grant, *2021*
Awarded around half a million dollars to hire engineers and interns for research on aligning language models with human preferences.

Open Philanthropy AI Fellowship Recipient, *2020*
One of 10 PhD students chosen for fellowship support for work on improving the long-term impact of AI.

New York Academy of Sciences, STAR Talk Winner, *2020*
Awarded for work on "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks."

NSF Graduate Research Fellowship Recipient, *2018*
One of 2,000 recipients chosen (out of 12,000).

Hertz Fellowship Finalist, *2018*
One of 40 finalists selected (out of 700).

Rice Engineering Department Outstanding Senior Award, *2018*
Awarded to one senior in the Rice Engineering department for academic and creative achievement.

Rice Computer Science Junior Merit Award, *2017*
Awarded to one junior in Computer Science for academic and creative achievement.

Kleiner Perkins Caufield and Byers Fellowship, *2016*
One of 54 KPCB Engineering Fellows selected (out of 2,500).

Chevron Computer Science Scholarship Winner, *2015*
Awarded to one Rice University freshman in Computer Science.

American Mathematics Competition: 90th Nationwide, *2012*
Amongst 10th graders in the U.S. (Certificate of Distinction).

Teaching, Service, and Activities

African Master's of Machine Intelligence Lab Lecturer, *2021*

Stanford Existential Risk Initiative: Supervisor
for project on learning to generate advice using human feedback, *2021*

Reviewer for EMNLP, NeurIPS, ICLR, CoNLL, ACL, ICML, *2018-present*

NeurIPS Visually-Grounded Interaction and Language Workshop Organizer, *2018*

NYU: Guest Lecture for Computer Vision: "RNNs and Image Captioning," *Fall 2018*

Workshop on AI Safety, Machine Intelligence Research Institute, *2017*

Rice University: Teaching Assistant for Computational Thinking, *Fall 2016*

Rice University: Teaching Assistant for Algorithms and Discrete Math, *Spring 2016*

St. Mark's Catholic School: MathCounts Program Founder & Coach, *2012–2014*