Ethan Perez | CV

✉ ethan@anthropic.com • 🌐 ethanperez.net • ethanjperez

Current Positions

Anthropic	**Senior Staff Research Scientist; Adversarial Robustness Team Lead**
AI Alignment Science Team	*May 2022–*
FAR AI	**Co-founder; Research Advisor**
	October 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

"Rapid Response: Mitigating LLM Jailbreaks with a Few Examples" *arXiv 2024*. Alwin Peng, Julian Michael, Henry Sleight, **Ethan Perez**, Mrinank Sharma.

"Sabotage Evaluations for Frontier Models" *arXiv 2024*. Joe Benton, Misha Wagner, Eric Christiansen, Cem Anil, **Ethan Perez**, Jai Srivastav, Esin Durmus, Deep Ganguli, Shauna Kravec, Buck Shlegeris, Jared Kaplan, Holden Karnofsky, Evan Hubinger, Roger Grosse, Samuel R Bowman, David Duvenaud.

"Looking Inward: Language Models Can Learn About Themselves by Introspection" *arXiv 2024*. Felix J Binder*, James Chua*, Tomek Korbak, Henry Sleight, John Hughes, Robert Long, **Ethan Perez**, Miles Turpin, Owain Evans.

"Language Models Learn to Mislead Humans via RLHF" *arXiv 2024*.
Jiaxin Wen, Ruiqi Zhong, Akbir Khan, **Ethan Perez**, Dylan Jacob Steinhardt, Minlie Huang, Samuel R Bowman, He He, Shi Feng.

"Latent Adversarial Training Improves Robustness to Persistent Harmful Behaviors in LLMs" *arXiv 2024*. Abhay Sheshadri*, Aidan Ewart*, Phillip Guo*, Aengus Lynch*, Cindy Wu*, Vivek Hebbar*, Henry Sleight, Asa Cooper Stickland, **Ethan Perez**, Dylan Hadfield-Menell, Stephen Casper.

"When Do Universal Image Jailbreaks Transfer Between Vision-Language Models?" *arXiv 2024*.
Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristóbal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, John Hughes, Rajashree Agrawal, Mrinank Sharma, Scott Emmons, Sanmi Koyejo, **Ethan Perez**.

"Sycophancy to subterfuge: Investigating reward-tampering in large language models" *arXiv 2024*.
Carson Denison*, Monte MacDiarmid, Fazl Barez, David Duvenaud, Shauna Kravec, Samuel Marks, Nicholas Schiefer, Ryan Soklaski, Alex Tamkin, Jared Kaplan, Buck Shlegeris, Samuel R Bowman, **Ethan Perez**, Evan Hubinger*.

"Many-shot jailbreaking" *NeurIPS 2024*.
Cem Anil, Esin Durmus, Nina Rimsky, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson,

Meg Tong, Jesse Mu, Daniel J Ford, Francesco Mosconi, Rajashree Agrawal, Rylan Schaeffer, Naomi Bashkansky, Samuel Svenningsen, Mike Lambert, Ansh Radhakrishnan, Carson Denison, Evan J Hubinger, Yuntao Bai, Trenton Bricken, Timothy Maxwell, Nicholas Schiefer, James Sully, Alex Tamkin, Tamera Lanham, Karina Nguyen, Tomasz Korbak, Jared Kaplan, Deep Ganguli, Samuel R Bowman, **Ethan Perez***, Roger Grosse*, David Duvenaud*.

"Bias-Augmented Consistency Training Reduces Biased Reasoning in Chain-of-Thought" *arXiv 2024*.
James Chua*, Edward Rees*, Hunar Batra, Samuel R Bowman, Julian Michael, **Ethan Perez**, Miles Turpin.

"Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting" *NeurIPS 2024*.
Miles Turpin, Julian Michael, **Ethan Perez**, Samuel Bowman.

"Debating with More Persuasive LLMs Leads to More Truthful Answers" ***ICML 2024 Best Paper Award***.
Akbir Khan*, John Hughes*, Dan Valentine*, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktaschel, **Ethan Perez**.

"Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training" *arXiv 2024*.
Evan Hubinger*, Carson Denison*, Jesse Mu*, Mike Lambert*, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, ..., Nicholas Schiefer*, **Ethan Perez***.

'Learning from Natural Language Feedback" *2024*.
Angelica Chen*, Jérémy Scheurer*, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, **Ethan Perez**

"Towards Evaluating AI Systems for Moral Status Using Self-Reports" *arXiv 2023*.
Ethan Perez, Robert Long.

"Specific versus General Principles for Constitutional AI" *arXiv 2023*.
Sandipan Kundu*, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, **Ethan Perez**, ..., Sam McCandlish, Jared Kaplan*.

"Towards Understanding Sycophancy in Language Models" *arXiv 2023*.
Mrinank Sharma*, Meg Tong*, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, ..., Miranda Zhang, **Ethan Perez**.

"Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning" *arXiv 2023*.
Juan Rocamonde, Victoriano Montesinos, Elvis Nava, **Ethan Perez***, David Lindner*.

"Studying Large Language Model Generalization with Influence Functions" *arXiv 2023*.
Roger Grosse*, Juhan Bae*, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, **Ethan Perez**, ..., Jared Kaplan, Samuel R. Bowman.

"Measuring Faithfulness in Chain-of-Thought Reasoning" *arXiv 2023*.
Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, ..., Samuel R Bowman, **Ethan Perez**.

"Question Decomposition Improves the Faithfulness of Model-Generated Reasoning" *arXiv 2023*.
Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, ..., Samuel R Bowman, **Ethan Perez**.

"Inverse Scaling: When Bigger Isn't Better" *arXiv 2023*.
Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan

McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, ..., Samuel R. Bowman, **Ethan Perez**.

“Language Models Don’t Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting” *arXiv 2023*.
Miles Turpin, Julian Michael, **Ethan Perez**, Samuel R Bowman.

“Training Language Models with Language Feedback at Scale” *arXiv 2023*.
Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, **Ethan Perez**.

“Improving Code Generation by Training with Natural Language Feedback” *arXiv 2023*.
Angelica Chen, Jérémy Scheurer, Tomasz Korbak, Jon Ander Campos, Jun Shern Chan, Samuel R Bowman, Kyunghyun Cho, **Ethan Perez**.

“Pretraining Language Models with Human Preferences” *ICML 2023*.
Tomasz Korbak, Kejian Shi, Angelica Chen, Rasika Bhalerao, Christopher L Buckley, Jason Phang, Samuel R Bowman, **Ethan Perez**.

“The Capacity for Moral Self-Correction in Large Language Models” *arXiv 2022*.
Deep Ganguli*, Amanda Askell*, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, Dawn Drain, Dustin Li, Eli Tran-Johnson, **Ethan Perez**, ..., Samuel R Bowman, Jared Kaplan.

“Discovering Language Model Behaviors with Model-Written Evaluations” *arXiv 2022*.
Ethan Perez, Sam Ringer*, Kamilė Lukošiūtė*, Karina Nguyen*, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, ..., Nicholas Schiefer, Jared Kaplan.

“Constitutional AI: Harmlessness from AI Feedback” *arXiv 2022*.
Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, **Ethan Perez**, ..., Tom Brown, Jared Kaplan.

“Measuring Progress on Scalable Oversight for Large Language Models” *arXiv 2022*.
Samuel R Bowman, Jeeyoon Hyun, **Ethan Perez**, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, ..., Ben Mann, Jared Kaplan.

“Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned” *arXiv 2022*.
Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, **Ethan Perez**, ..., Jared Kaplan, Jack Clark.

“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

New York University — **Research Advisor**
Alignment Research Group — *September 2022-May 2024*

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

"Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting."
CHAI 2023, May 2023.

"Discovering AI RIsks with AI."
EA Global: Bay Area, February 2023.

"Discovering Language Model Behaviors with Model-Written Evaluations."
IBM Research Seminar, February 2023.

"Discovering Language Model Behaviors with Model-Written Evaluations."
AI Safety Workshop in Orinda, December 2022.

"Aligning Language Models with Human Preferences."
Rice University CS Fall Colloquium, September 2022.

"Aligning Language Models with Human Preferences."
IICCSSS 2022, September 2022.

"Aligning Language Models with Human Preferences."
CHAI 2022, June 2022.

"Aligning Language Models with Human Preferences."
Bay Area NLP, April 2022.

"Aligning Language Models with Human Preferences."

Aleph Alpha Talks, March 2022.

"True Few-shot Learning with Language Models."
SKKU AI Workshop, October 2021.

"Rissanen Data Analysis."
2021 2nd Mila-NYU-Samsung workshop, October 2021.

"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, *2021–2024*
Awarded around 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

Stanford Existential Risk Initiative: Supervisor
Supervising 8 interns on projects related to aligning language models with human preferences, *2023*

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*