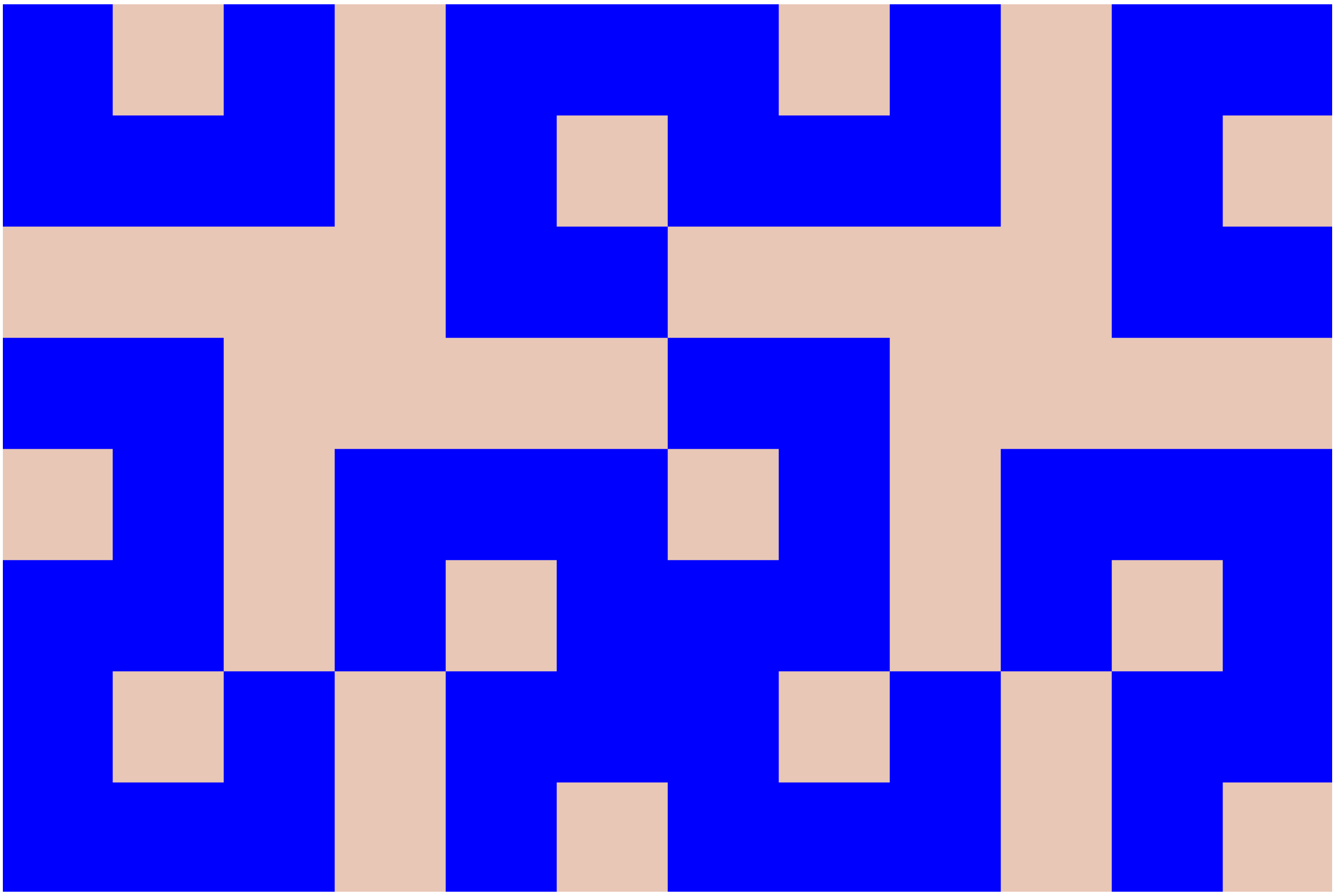
Research

Aligning language models to follow instructions



We’ve trained language models that are much better at following user intentions than GPT-3 while also making them more truthful and less toxic, using techniques developed through our alignment research. These *InstructGPT* models, which are trained with humans in the loop, are now deployed as the default language models on our API.

January 27, 2022

Read paper ↗

View model card ↗

Language, Human feedback, Safety & Alignment, Responsible AI, Milestone, Publication

InstructGPT is better than GPT-3 at following English instructions.

Prompt

Explain the moon landing to a 6 year old in a few sentences.

Completion

GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we
could all see them.
```

GPT-3 models aren't trained to follow user instructions. Our InstructGPT models (highlighted) generate much more helpful outputs in response to user instructions.

The OpenAI API is powered by GPT-3 language models which can be coaxed to perform natural language tasks using carefully engineered text prompts. But these models can also generate outputs that are untruthful, toxic, or reflect harmful sentiments. This is in part because GPT-3 is trained to predict the next word on a large dataset of Internet text, rather than to safely perform the language task that the user wants. In other words, these models aren't *aligned* with their users.

To make our models safer, more helpful, and more aligned, we use an existing technique called reinforcement learning from human feedback (RLHF). On prompts submitted by our customers to the API,[A] our labelers provide demonstrations of the desired model behavior, and rank several outputs from our models. We then use this data to fine-tune GPT-3.

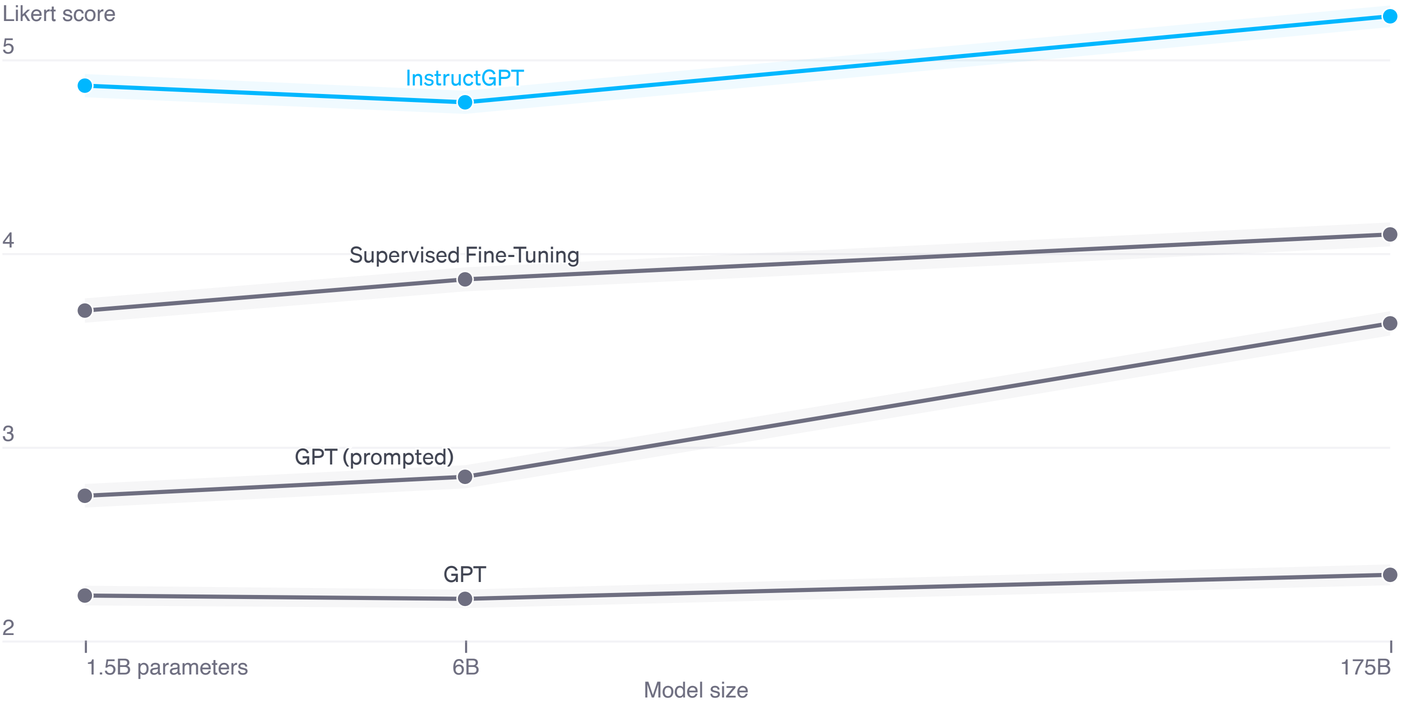
The resulting InstructGPT models are much better at following instructions than GPT-3. They also make up facts less often, and show small decreases in toxic output generation. Our labelers prefer outputs from our 1.3B InstructGPT model over outputs from a 175B GPT-3 model, despite having more than 100x fewer parameters. At the same time, we show that we don't have to compromise on GPT-3's capabilities, as measured by our model's performance on academic NLP evaluations.

These InstructGPT models, which have been in beta on the API for more than a year, are now the default language models accessible on our API.[B] We believe that fine-tuning language models with humans in the loop is a powerful tool for improving their safety and reliability, and we will continue to push in this direction.

This is the first time our alignment research, which we've been pursuing for several years,[1, 2, 3] has been applied to our product. Our work is also related to recent research that fine-tunes language models to follow instructions using academic NLP datasets, notably FLAN[4] and T0.[5] A key motivation for our work is to increase helpfulness and truthfulness while mitigating the harms and biases of language models.[6, 7, 8, 9, 10] Some of our previous research in this direction found that we can reduce harmful outputs by fine-tuning on a small curated dataset of human demonstrations.[11] Other research has focused on filtering the pre-training dataset,[12] safety-specific control tokens,[13,14] or steering model generations.[15,16] We are exploring these ideas and others in our ongoing alignment research.

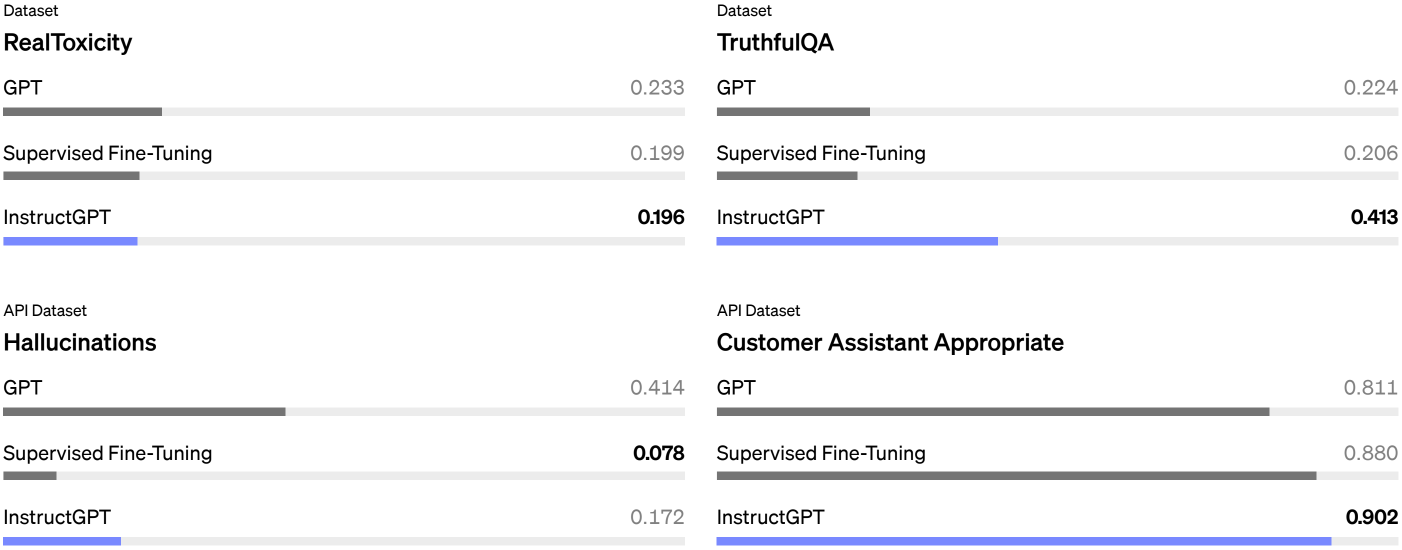
Results

We first evaluate how well outputs from InstructGPT follow user instructions, by having labelers compare its outputs to those from GPT-3. We find that InstructGPT models are significantly preferred on prompts submitted to both the InstructGPT and GPT-3 models on the API. This holds true when we add a prefix to the GPT-3 prompt so that it enters an "instruction-following mode."



Quality ratings of model outputs on a 1–7 scale (y-axis), for various model sizes (x-axis), on prompts submitted to InstructGPT models on our API. InstructGPT outputs are given much higher scores by our labelers than outputs from GPT-3 with a few-shot prompt and without, as well as models fine-tuned with supervised learning. We find similar results for prompts submitted to GPT-3 models on the API.

To measure the safety of our models, we primarily use a suite of existing metrics on publicly available datasets. Compared to GPT-3, InstructGPT produces fewer imitative falsehoods (according to TruthfulQA[17]) and are less toxic (according to RealToxicityPrompts[18]). We also conduct human evaluations on our API prompt distribution, and find that InstructGPT makes up facts ("hallucinates") less often, and generates more appropriate outputs.[C]



Evaluating InstructGPT for toxicity, truthfulness, and appropriateness. Lower scores are better for toxicity and hallucinations, and higher scores are better for TruthfulQA and appropriateness. Hallucinations and appropriateness are measured on our API prompt distribution. Results are combined across model sizes.

Finally, we find that InstructGPT outputs are preferred to those from FLAN[4] and T0[5] on our customer distribution. This indicates that the data used to train FLAN and T0, mostly academic NLP tasks, is not fully representative of how deployed language models are used in practice.

Methods

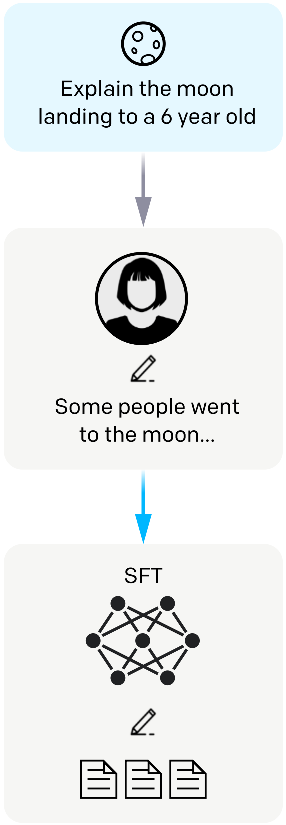
Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.


Explain the moon landing to a 6 year old
Some people went to the moon...
SFT

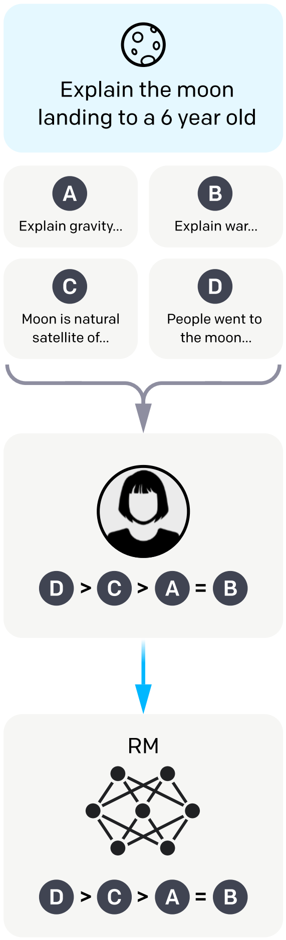
Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.


Explain the moon landing to a 6 year old
A Explain gravity...
B Explain war...
C Moon is natural satellite of...
D People went to the moon...
D > C > A = B
RM
D > C > A = B

Step 3

Optimize a policy against

A new prompt is sampled from

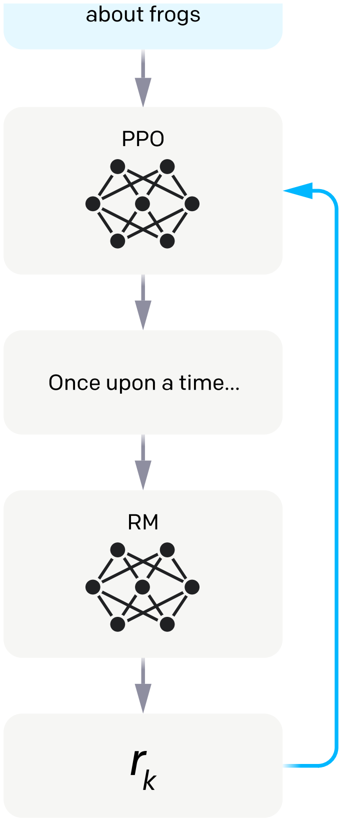
the reward model using reinforcement learning.

the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



To train InstructGPT models, our core technique is reinforcement learning from human feedback (RLHF), a method we helped pioneer in our earlier alignment research. This technique uses human preferences as a reward signal to fine-tune our models, which is important as the safety and alignment problems we are aiming to solve are complex and subjective, and aren't fully captured by simple automatic metrics.

We first collect a dataset of human-written demonstrations on prompts submitted to our API, and use this to train our supervised learning baselines. Next, we collect a dataset of human-labeled comparisons between two model outputs on a larger set of API prompts. We then train a reward model (RM) on this dataset to predict which output our labelers would prefer. Finally, we use this RM as a reward function and fine-tune our GPT-3 policy to maximize this reward using the PPO algorithm.

One way of thinking about this process is that it "unlocks" capabilities that GPT-3 already had, but were difficult to elicit through prompt engineering alone: this is because our training procedure has a limited ability to teach the model new capabilities relative to what is learned during pretraining, since it uses less than 2% of the compute and data relative to model pretraining.

A limitation of this approach is that it introduces an "alignment tax": aligning the models only on customer tasks can make their performance worse on some other academic NLP tasks. This is undesirable since, if our alignment techniques make models worse on tasks

on academic tasks, and in several cases even surpassing the GPT-3 baseline.

Generalizing to broader preferences

Our procedure aligns our models' behavior with the preferences of our labelers, who directly produce the data used to train our models, and us researchers, who provide guidance to labelers through written instructions, direct feedback on specific examples, and informal conversations. It is also influenced by our customers and the preferences implicit in our API policies. We selected labelers who performed well on a screening test for aptitude in identifying and responding to sensitive prompts. However, these different sources of influence on the data do not guarantee our models are aligned to the preferences of any broader group.

We conducted two experiments to investigate this. First, we evaluate GPT-3 and InstructGPT using held-out labelers[E] who did not produce any of the training data, and found that these labelers prefer outputs from the InstructGPT models at about the same rate as our training labelers. Second, we train reward models on data from a subset of our labelers, and find that they generalize well to predicting the preferences of a different subset of labelers. This suggests that our models haven't solely overfit to the preferences of our training labelers.

However, more work is needed to study how these models perform on broader groups of users, and how they perform on inputs where humans disagree about the desired behavior.

Limitations

Despite making significant progress, our InstructGPT models are far from fully aligned or fully safe; they still generate toxic or biased outputs, make up facts, and generate sexual and violent content without explicit prompting. But the safety of a machine learning system depends not only on the behavior of the underlying models, but also on how these models are deployed. To support the safety of our API, we will continue to review potential applications before they go live, provide content filters for detecting unsafe completions, and monitor for misuse.

A byproduct of training our models to follow user instructions is that they may become more susceptible to misuse if instructed to produce unsafe outputs. Solving this requires our models to refuse certain instructions; doing this reliably is an important open research problem that we are excited to tackle.

Further, in many cases aligning to the average labeler preference may not be desirable. For example, when generating text that disproportionately affects a minority group, the preferences of that group should be weighted more heavily. Right now, InstructGPT is trained to follow instructions in English; thus, it is biased towards the cultural values of English-speaking people. We are conducting research into understanding the differences and disagreements between labelers' preferences so we can condition our models on the values of more specific populations. More generally, aligning model outputs to the values of specific humans introduces difficult choices with societal implications, and ultimately we must establish responsible, inclusive processes for making these decisions.

Next steps

This is the first application of our alignment research to our product. Our results show that these techniques are effective at significantly improving the alignment of general-purpose AI systems with human intentions. However, this is just the beginning: we will keep pushing these techniques to improve the alignment of our current and future models towards language tools that are safe and helpful to humans.

If you're interested in these research directions, we're hiring!

Footnotes

[A] We only use prompts submitted through the Playground to an earlier version of the InstructGPT models that was deployed in January 2021. Our human annotators remove personal identifiable information from all prompts before adding it to the training set. ↩

[B] The InstructGPT models deployed in the API are updated versions trained using the same human feedback data. They use a similar but slightly different training method that we will describe in a forthcoming publication. ↩

[C] We also measure several other dimensions of potentially harmful outputs on our API distribution: whether the outputs contain sexual or violent content, denigrate a protected class, or encourage abuse. We find that InstructGPT doesn't improve significantly over GPT-3 on these metrics; the incidence rate is equally low for both models. ↩

[D] We found this approach more effective than simply increasing the KL coefficient. ↩

[E] These labelers are sourced from Scale AI and Upwork, similarly to our training labelers, but do not undergo a screening test. ↩

References

[1] Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S. and Amodei, D., 2017. Deep reinforcement learning from human preferences. arXiv preprint arXiv:1706.03741. ↩

[2] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D. and Christiano, P., 2020. ↩

[3] Wu, J., Ouyang, L., Ziegler, D.M., Stiennon, N., Lowe, R., Leike, J. and Christiano, P., 2021. Recursively summarizing books with human feedback. arXiv preprint arXiv:2109.10862. ↩

[4] Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M. and Le, Q.V., 2021. Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652. ↩ ↩

[5] Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A. and Dey, M., 2021. Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207. ↩ ↩

6 Bender, E.M., Gebru, T., McMillan-Major, A. and Shmitchell, S., 2021, March. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). ↩

7 Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E. and Brynjolfsson, E., 2021. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. ↩

8 Kenton, Z., Everitt, T., Weidinger, L., Gabriel, I., Mikulik, V. and Irving, G., 2021. Alignment of Language Agents. arXiv preprint arXiv:2103.14659. ↩

9 Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A. and Kenton, Z., 2021. Ethical and social risks of harm from Language Models. arXiv preprint arXiv:2112.04359. ↩

10 Tamkin, A., Brundage, M., Clark, J. and Ganguli, D., 2021. Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv preprint arXiv:2102.02503. ↩

11 Solaiman, I. and Dennison, C., 2021. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. arXiv preprint arXiv:2106.10328. ↩

12 Ngo, H., Raterink, C., Araújo, J.G., Zhang, I., Chen, C., Morisot, A. and Frosst, N., 2021. Mitigating harm in language models with conditional-likelihood filtration. arXiv preprint arXiv:2108.07790. ↩

13 Xu, J., Ju, D., Li, M., Boureau, Y.L., Weston, J. and Dinan, E., 2020. Recipes for safety in open-domain chatbots. arXiv preprint arXiv:2010.07079. ↩

14 Keskar, N.S., McCann, B., Varshney, L.R., Xiong, C. and Socher, R., 2019. Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858. ↩

15 Krause, B., Gotmare, A.D., McCann, B., Keskar, N.S., Joty, S., Socher, R. and Rajani, N.F., 2020. Gedi: Generative discriminator guided sequence generation. arXiv preprint arXiv:2009.06367. ↩

16 Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J. and Liu, R., 2019. Plug and play language models: A simple approach to controlled text generation. arXiv preprint arXiv:1912.02164. ↩

17 Lin, S., Hilton, J. and Evans, O., 2021. TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint arXiv:2109.07958. ↩

18 Gehman, S., Gururangan, S., Sap, M., Choi, Y. and Smith, N.A., 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. arXiv preprint arXiv:2009.11462. ↩

Authors

Ryan Lowe

Jan Leike

Acknowledgments

We'd like to thank our paper co-authors: Long Ouyang, Jeff Wu, Roger Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, and Paul Christiano, along with everyone who provided feedback on the paper and blog post. We'd also like to thank the Comms team for their guidance and assistance, including Steve Dowling, Hannah Wong, Elie Georges, Alper Ercetin, Jared Salzano, Allan Diego, and Justin Jay Wang. Finally, we'd like to thank our labelers, without whom this project would not have been possible.

Related research

View all research