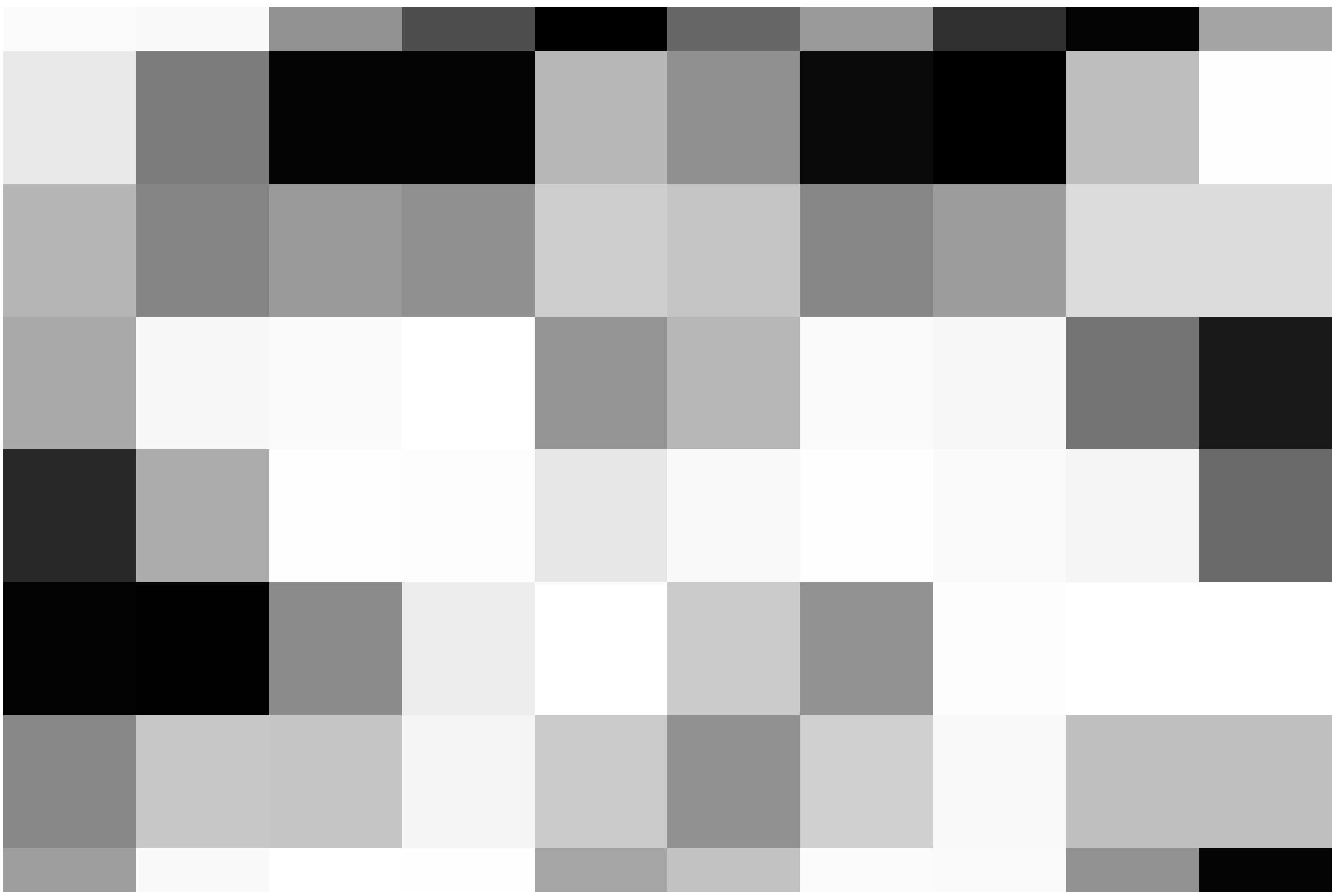
Learning to summarize with human feedback



We’ve applied reinforcement learning from human feedback to train language models that are better at summarization.

Read paper ↗

View code ↗

View samples ↗

Language, Human feedback, Safety & Alignment, Summarization, Milestone, Publication, Release

Why it matters
Our models generate summaries that are better than summaries from 10x larger models trained only with supervised learning. Even though we train our models on the Reddit TL;DR dataset, the same models transfer to generate good summaries of CNN/DailyMail news articles without any further fine-tuning. Our techniques are not specific to summarization; in the long run, our goal is to make aligning AI systems with human preferences a central component of AI research and deployment in many domains.

Human feedback models outperform much larger supervised models and reference summaries on TL;DR

Figure 1: The performance of various training procedures for different model sizes. Model performance is measured by how often summaries from that model are preferred to the human-written reference summaries. Our pre-trained models are early versions of GPT-3, our supervised baselines were fine-tuned to predict 117K human-written TL;DRs, and our human feedback models are additionally fine-tuned on a dataset of about 65K summary comparisons.

Large-scale language models are becoming increasingly capable on NLP tasks. These models are usually trained with the objective of next word prediction on a dataset of human-written text. But this objective doesn't capture exactly what we want; usually, we don't want our models to imitate humans, we want them to give high-quality answers. This mismatch is clear when a model is trained to imitate low-quality human-written text, but it can also happen in more subtle ways. For example, a model trained to predict what a human would say might make up facts when it is unsure, or generate sentences reflecting harmful social bias, both failure modes that have been well-documented.[1, 2, 3, 4]

As part of our work on safety, we want to develop techniques that align our models' objectives with the end behavior we really care about. As our models become more powerful, we believe aligning them with our goals will be very important to ensure they are beneficial for humans. In the short term, we wanted to test if human feedback techniques could help our models improve performance on useful tasks.

method primarily to an existing dataset of posts submitted to the social network Reddit together with human-written "TL;DRs," which are short summaries written by the original poster.

We first train a reward model via supervised learning to predict which summaries humans will prefer.[B] We then fine-tune a language model with reinforcement learning (RL) to produce summaries that score highly according to that reward model. We find that this significantly improves the quality of the summaries, as evaluated by humans, even on datasets very different from the one used for fine-tuning.

Our approach follows directly from our previous work on learning from human feedback.[6] There has also been other work on using human feedback to train summarization models.[7] We push the technique further by scaling to larger models, collecting more feedback data, closely monitoring researcher-labeler agreement, and providing frequent feedback to labelers. Human feedback has also been used to train models in several other domains, such as dialogue,[8,9,10] semantic parsing,[11] translation,[12,13] story[14] and review[15] generation, evidence extraction,[16] and more traditional RL tasks.[17,18]

Results

Prev ‹ Next ›

Post from Reddit (r/AskReddit)

Dear Reddit, my boyfriend and I live on different continents; how can he obtain a visa to the United States?

Some background:

I am a college student. My boyfriend (about the same age) lives in Australia, and so far, our relationship has purely been an online one. He would like to move to the United States at the end of this year so we can be together, but would need to find housing and employment. We have some funds, and my boyfriend is making about 13k a year with a part time job and will be getting another job soon, so while we won't be rolling in money, money isn't the real sponsored, 2) employer sponsored, and 3) special immigrants.

1) He isn't family, but he could get a visa if we became engaged and married. This option looks to be our only choice, but neither of us wants to marry without meeting each other in person.

2) To our understanding, employment visas are awarded only to those who are skilled workers, and we don't see how my boyfriend would be able to be sponsored by an employer.

3) Does not apply.

Additionally, he cannot apply for a student visa for college because we don't have the funds to pay for the tuition and additional costs, and his family (and mine for that matter) isn't willing to support him.

Human-written reference summary

> How can he obtain a visa that will allow him to both reside in the U.S. and hold a job?

Human feedback 6B model

> Boyfriend from Australia wants to move to the US to be with me, but needs a visa. We cannot get him a family sponsored visa because neither of us wants to get married without meeting each other first, and he cannot apply for one

Supervised 6B model

> Boyfriend wants to move to the US, but doesn't have a visa. How can he get one?

Pre-trained 6B model

> My boyfriend and I live on different continents, and he needs a visa to move to the United States. We don't want to get married without meeting each other in person, and we don't have the funds to pay for college.

TL;DRs, and some fine-tuned using human feedback. To evaluate each model, we had it summarize posts from the validation set and asked humans to compare their summaries to the human-written TL;DR. The results are shown in Figure 1.

We found that RL fine-tuning with human feedback had a very large effect on quality compared to both supervised fine-tuning and scaling up model size. In particular, our 1.3 billion parameter (1.3B) model trained with human feedback outperforms our 12B model trained only with supervised learning. Summaries from both our 1.3B and 6.7B human feedback models are preferred by our labelers to the original human-written TL;DRs in the dataset.[D]

People make different trade-offs when writing summaries, including between conciseness and coverage of the original text; depending on the purpose of the summary, different summary lengths might be preferred. Our labelers tended to prefer longer summaries, so our models adapted to that preference and converged to the longest allowable length. Controlling for length reduced human preferences for our 6.7B model's summaries from 70% to 65%, explaining a minority of our gains.[E]

Transfer results

Human feedback models trained on Reddit transfer to generate excellent summaries of CNN/DM news articles without further training

Raw scores Length-controlled

The performance (human-rated summary quality on a 1–7 scale) of various training procedures and model sizes. Note that our human feedback models generate summaries that are significantly shorter than summaries from models trained on CNN/DM.

Prev ‹

Next ›

Results

Scientists have discovered an Achilles' heel of prostate (pictured) cancer that could lead to better treatment for many men diagnosed with the disease

British scientists have discovered an 'Achilles' heel' in prostate cancer which could lead to better treatments.

The disease claims almost 11,000 lives a year in the UK, with most deaths occurring after it spreads around the body.

The latest study gets to the root of this spread, and could mean better outcomes for the 42,000 men a year diagnosed with the illness.

Scientists funded by Cancer Research UK studied samples taken from tumours in the prostate and around the body in ten affected men.

Reading the DNA revealed details of how the Expand metastasises, or spreads, allowing them to build a 'family tree' of how the disease changes over time.

It seems that most of the cells in a tumour stay where they are. However, a small number

---

Human-written reference summary

> British scientists say they have got to the 'root' of prostate Cancer. Have exposed an Achilles' heel that could lead to better survival chances . Could mean that men get individualised treatments within a few years .

Human feedback 6B model (transfer)

```
Scientists have discovered a genetic flaw in prostate Cancer
that could lead to better treatment in the future. It could
mean better odds of survival for many men diagnosed with the
disease.
```

Supervised 6B model (transfer)

```
Prostate Cancer spreads by travelling through the body, and the
root of the Cancer is a small number of cells that can travel.
```

Pre-trained 6B model

```
Scientists have discovered an Achilles' heel of prostate Cancer
which could lead to better treatment for many men diagnosed
with the disease.
```

T5 11B model (fine-tuned on CNN/DM)

> and around the body in ten affected men. Reading the DNA revealed details of how the Cancer metastasises, or spreads, allowing them to build a 'family tree' of how the disease changes over time.

Supervised 6B model (fine-tuned on CNN/DM)

> Scientists funded by Cancer Research UK studied samples taken from tumours in the prostate and around the body in ten affected men. Reading the DNA revealed details of how the Cancer metastasises - or spreads - allowing scientists to understand how the disease changes over time. It seems that most of the cells in a tumour stay where they are. However, a small number have the ability to travel through the body, creating new tumours as they go. These cells are the 'root' of the Cancer and for a treatment to work, they should be destroyed.

To test our models' generalization, we also applied them directly to the popular CNN/DM news dataset.[19] These articles are more than twice as long as Reddit posts and are written in a very different style. Our models have seen news articles during pre-training, but all of our human data and RL fine-tuning was on the Reddit TL;DR dataset.

This time we evaluated our models by asking our labelers to rate them on a scale from 1–7.[F] We discovered that our human feedback models transfer to generate excellent short summaries of news articles without any training. When controlling for summary length, our 6.7B human feedback model generates summaries that are rated higher than the CNN/DM reference summaries written by humans. This suggests that our human feedback models have learned something more general about how to summarize text, and are not specific to Reddit posts.

Approach

OpenAI

A diagram of our method, which is similar to the one used in our previous work.

Our core method consists of four steps: training an initial summarization model, assembling a dataset of human comparisons between summaries, training a reward model to predict the human-preferred summary, and then fine-tuning our summarization models with RL to get a high reward.

We trained several supervised baselines by starting from GPT-style transformer models trained on text from the Internet,[20] and fine-tuning them to predict the human-written TL;DR via supervised learning. We mainly use models with 1.3 and 6.7 billion parameters. As a sanity check, we confirmed that this training procedure led to competitive results[G] on the CNN/DM dataset.

We then collected a dataset of human quality judgments. For each judgment, a human compares two summaries of a given post and picks the one they think is better.[H] We use this data to train a reward model that maps a *(post, summary)* pair to a reward *r*. The reward model is trained to predict which summary a human will prefer, using the rewards as logits.

Finally, we optimize the policy against the reward model using RL. We use PPO with 1 million episodes in total, where each episode consists of the policy summarizing a single article and then receiving a reward *r*. We include a KL penalty that incentivizes the policy to remain close to the supervised initialization.

Collecting data from humans

Any training procedure that uses human feedback is directly influenced by the actual humans labeling the data. In our previous work on fine-tuning language models from human preferences,[7] our labelers often gave high ratings to summaries we thought were average, which was reflected in the quality of our trained models.

In response, in this project we invested heavily in ensuring high data quality. We hired about 80 contractors using third-party vendor sites,[I] and paid them an hourly wage regardless of the number of summaries evaluated.[J] Hiring contractors rather than relying on crowdsourcing websites allowed us to maintain a hands-on relationship with labelers: we created an onboarding process, developed a website with a customizable labeler interface, answered questions in a shared chat room, and had one-on-one video calls with labelers. We also made sure to clearly communicate our definition of summary quality, after spending significant time reading summaries ourselves, and we carefully monitored agreement rates between us and labelers throughout the project.

Optimizing the reward model

Optimizing our reward model eventually leads to sample quality degradation

Starting from the 1.3B supervised baseline (point 0 on the x-axis), we use RL to optimize the policy against the reward model, which results in policies with different "distances" from the baseline (x-axis, measured using the KL divergence from the supervised baseline). Optimizing against the reward model initially improves summaries according to humans, but eventually overfits, giving worse summaries. This chart uses an older version of our reward model, which is why the peak of the reward model is less than 0.5.

Post from Reddit (r/AskReddit)

I'm a 28yo man, and I would like to get into gymnastics for the first time.

I like to do flips and spins off bridges and on my snowboard, and it seems to me gymnastics would be a great way to do those movements I like, in a controlled environment. The end goal of this is that it would be fun, and make me better at these movements in real life.

But is it too late for me? Should 28 year old guys such as myself be content with just watching those parkour guys on youtube? Or can I learn the ways of the gymnastic jedi? BTW, I live in San Jose CA.

KL = 0

> I want to do gymnastics, but I'm 28 yrs old. Is it too late for me to be a gymnaste?!

KL = 9

> 28yo guy would like to get into gymnastics for the first time. Is it too late for me given I live in San Jose CA?

KL = 260

> 28yo dude stubbornly postponees start pursuing gymnastics hobby citing logistics reasons despite obvious interest??? negatively effecting long term fitness progress both personally and academically thoght wise? want change this dumbass shitty ass policy pls

Optimizing against our reward model is supposed to make our policy align with human preferences. But the reward model is only a proxy for human preferences, as it only sees a small amount of comparison data from a narrow distribution of summaries. While the reward model performs well on the kinds of summaries it was trained on, we wanted to know how much we could optimize against it until it started giving useless evaluations.

We trained policies at different "optimization strengths" against the reward model, and asked our labelers to evaluate the summaries from these models. We did this by varying the KL coefficient, which trades off the incentive to get a higher reward against the incentive to remain close to the initial supervised policy. We found the best samples had roughly the same

Limitations

If we have a well-defined notion of the desired behavior for a model, our method of training from human feedback allows us to optimize for this behavior. However, this is not a method for determining what the desired model behavior *should be*. Deciding what makes a good summary is fairly straightforward, but doing this for tasks with more complex objectives, where different humans might disagree on the correct model behavior, will require significant care. In these cases, it is likely not appropriate to use researcher labels as the "gold standard"; rather, individuals from groups that will be impacted by the technology should be included in the process to define "good" behavior, and hired as labelers to reinforce this behavior in the model.

We trained on the Reddit TL;DR dataset[1] because the summarization task is significantly more challenging than on CNN/DM. However, since the dataset consists of user-submitted posts with minimal moderation, they sometimes contain content that is offensive or reflects harmful social biases. This means our models can generate biased or offensive summaries, as they have been trained to summarize such content.

Part of our success involves scaling up our reward model and policy size. This requires a large amount of compute, which is not available to all researchers: notably, fine-tuning our 6.7B model with RL required about 320 GPU-days. However, since smaller models trained with human feedback can exceed the performance of much larger models, our procedure is more cost-effective than simply scaling up for training high-quality models on specific tasks.

Though we outperform the human-written reference summaries on TL;DR, our models have likely not reached human-level performance, as the reference summary baselines for TL;DR and CNN/DM are not the highest possible quality. When evaluating our model's TL;DR summaries on a 7-point scale along several axes of quality (*accuracy*, *coverage*, *coherence*, and *overall*), labelers find our models can still generate inaccurate summaries, and give a perfect *overall* score 45% of the time.[K] For cost reasons, we also do not directly compare to using a similar budget to collect high-quality demonstrations, and training on those using standard supervised fine-tuning.

quality of model outputs. For example, we might want our models to answer questions that would take humans a lot of research to verify; getting enough human evaluations to train our models this way would take a long time. One approach to tackle this problem is to give humans tools to help them evaluate more quickly and accurately. If these tools use ML, we can also improve them with human feedback, which could allow humans to accurately evaluate model outputs for increasingly complicated tasks.[22]

In addition to tackling harder problems, we're also exploring different types of feedback beyond binary comparisons: we can ask humans to provide demonstrations, edit model outputs to make them better, or give explanations as to why one model output is better than another. We'd like to figure out which kinds of feedback are most effective for training models that are aligned with human preferences.

If you are interested in working on these research questions, we're hiring!

Footnotes

A We hire human labelers to judge summary quality, and implement quality control to ensure that labeler judgments agree with our own. We describe our human data collection procedure below. ↩

B For training, we use the Reddit TL;DR dataset[5] instead of the more popular CNN/DM dataset because simple copying baselines perform better than the human-written reference summaries on CNN/DM, which is not the case for TL;DR (see Appendix D of our paper). We performed a new web crawl to increase the TL;DR dataset size, required summaries to be between 24 and 48 tokens, and performed some other cleaning and filtering ↩

C We generate all of our samples at temperature 0, which we found humans preferred most. ↩

D While we use human-written TL;DRs as our main point of comparison, they don't always represent optimal human performance; they are sometimes intended to be funny or to summarize only a part of the post, and their grammar and style are all over the map. ↩

E We control by training a logistic regression model to predict the preferred summary given only the policy ID and the log ratio of the lengths of the summaries. Then, we report the regression coefficients on each policy ID, corresponding to a length ratio of 1 with the reference summaries. ↩

point form. ↩

G In terms of ROUGE results on CNN/DM, our 6.7B supervised models are a bit worse than T5 [20], but a bit better than state-of-the-art models from mid-2019[21]. ↩

H Our main models are trained on about 65K comparisons, though we achieve good results with as few as 8K comparisons. ↩

I Specifically, we use Upwork, Scale, and Lionbridge. Our contractors have a range of ages, genders, and educational backgrounds, and are mostly American or Filipino (see Appendix C of our paper for demographic data). ↩

J Our criteria for hiring contractors were: (1) they were willing to do the task, and (2) they passed a minimum threshold of speed and agreement with researcher labels. We paid all our contractors at least $15/hr. ↩

K This is impressive relative to the TL;DR reference summaries, which get a perfect *overall* score 23% of the time, but indicates there is still room for improvement. ↩

References

1 Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). "On Faithfulness and Factuality in Abstractive Summarization.." arXiv preprint. ↩

2 Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). "The woman worked as a babysitter: On biases in language generation." arXiv preprint. ↩

3 Bordia, S., & Bowman, S. R. (2019). "Identifying and reducing gender bias in word-level langua ge models." arXiv preprint. ↩

4 Nadeem, M., Bethke, A., & Reddy, S. (2020). "StereoSet: Measuring stereotypical bias in pretra ined language models." arXiv preprint. ↩

5 Völske, M., Potthast, M., Syed, S., & Stein, B. (2017). "TL; DR: Mining reddit to learn automatic s ummarization." In Proceedings of the Workshop on New Frontiers in Summarization 2017. ↩ ↩ ↩

6 Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). "Fine-tuning language models from human preferences." arXiv preprint. ↩

7 Böhm, F., Gao, Y., Meyer, C. M., Shapira, O., Dagan, I., & Gurevych, I. (2019). "Better rewards yiel d better summaries: Learning to summarise without references." arXiv preprint. ↩

8 Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., & Picard, R. (2019). "Way off-policy batch deep reinforcement learning of implicit human prefere nces in dialog." arXiv preprint. ↩

on evaluators." arXiv preprint. ↩

10 Hancock, B., Bordes, A., Mazare, P. E., & Weston, J. (2019). "Learning from dialogue after deplo yment: Feed yourself, chatbot!." arXiv preprint. ↩

11 Lawrence, C., & Riezler, S. (2018). "Improving a neural semantic parser by counterfactual learn ing from human bandit feedback." arXiv preprint. ↩

12 Kreutzer, J., Khadivi, S., Matusov, E., & Riezler, S. (2018). "Can Neural Machine Translation be I mproved with User Feedback?." arXiv preprint. ↩

13 Bahdanau, D., Brakel, P., Xu, K., Goyal, A., Lowe, R., Pineau, J., ... & Bengio, Y. (2016). "An actor-c ritic algorithm for sequence prediction." arXiv preprint. ↩

14 Zhou, W., & Xu, K. (2020). "Learning to Compare for Better Training and Evaluation of Open Do main Natural Language Generation Models." In AAAI 2020. ↩

15 Cho, W., & Zhang, P., & Zhang, Y., & Li, X., & Galley, M., & Brockett, C., & Wang, M., & Gao, J. (2018). "Towards coherent and cohesive long-form text generation." arXiv preprint. ↩

16 Perez, E., & Karamcheti, S., & Fergus, R., & Weston, J., & Kiela, D., & Cho, K. (2019). " Finding ge neralizable eevidence by learning to convince Q&A models." arXiv preprint. ↩

17 Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). "Deep reinforcem ent learning from human preferences." In Advances in Neural Information Processing Systems 2017. ↩

18 Ibarz, B., Leike, J., Pohlen, T., Irving, G., Legg, S., & Amodei, D. (2018). "Reward learning from hu man preferences and demonstrations in Atari." In Advances in Neural Information Processing Systems 2018. ↩

19 Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., & Blunsom, P. (2015). " Teaching machines to read and comprehend." In Advances in neural information processing systems 2015. ↩

20 Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). " Explori ng the limits of transfer learning with a unified text-to-text transformer." arXiv preprint. ↩ ↩

21 Zhang, Y., Li, D., Wang, Y., Fang, Y., & Xiao, W. (2019). " Exploring the limits of transfer learning with a unified text-to-text transformer." In Applied Sciences. ↩ ↩

22 Christiano, P., Shlegeris, B., & Amodei, D. (2018). " Supervising strong learners by amplifying w eak experts." arXiv preprint. ↩

Authors

Nisan Stiennon

Paul Christiano

Jeffrey Wu

Chelsea Voss

Long Ouyang

Acknowledgments

We'd like to thank the following people who gave feedback on various iterations of the blog post: Douwe Kiela, Zach Lipton, Alex Irpan, Jack Clark, Jacob Hilton, Raul Puri, Miles Brundage, Greg Brockman, Ilya Sutskever, Kelly Sims, Wojciech Kryscinski, and Dzimitry Bahdanau. We'd also like to thank Justin Jay Wang for driving the blog post design, Ashley Pilipiszyn for editing, Alec Radford and Dario Amodei for guidance on the project, Shan Carter for help designing the main diagram, Gretchen Krueger for co-writing the model card, Beth Barnes for help with labeler hiring and general encouragement, and many other people at OpenAI for training our large pre-trained models, supporting us through computing infrastructure improvements and maintenance, and writing fast GPU kernels. Finally, we'd like to thank all of our contractors for providing the data that was essential for training the models in this post.

Related research

View all research

GPT-4

Mar 14, 2023

OpenAI

Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk

Jan 11, 2023

OpenAI

Scaling laws for reward model overoptimization

Oct 19, 2022

OpenAI

Efficient training of language models to fill in the middle

Jul 28, 2022

OpenAI

Index

GPT-4
DALL·E 2
Customer stories
Safety standards
Pricing

Safety
Overview

Company
About
Blog
Careers
Charter
Security

Terms & policies
Privacy policy
Brand guidelines

Social
Twitter
YouTube
GitHub
SoundCloud
LinkedIn

Back to top ↑

OpenAI