

[Try ChatGPT ↗](#) [Read more >](#)[API](#)[RESEARCH](#)[BLOG](#)[ABOUT](#)

# WebGPT: Improving the Factual Accuracy of Language Models through Web Browsing

December 16, 2021

5 minute read

We’ve fine-tuned GPT-3 to more accurately answer open-ended questions using a text-based web browser. Our prototype copies how humans research answers to questions online—it submits search queries, follows links, and scrolls up and down web pages. It is trained to cite its sources, which makes it easier to give feedback to improve factual accuracy. We’re excited about developing more truthful AI,<sup>1</sup> but challenges remain, such as coping with unfamiliar types of questions.

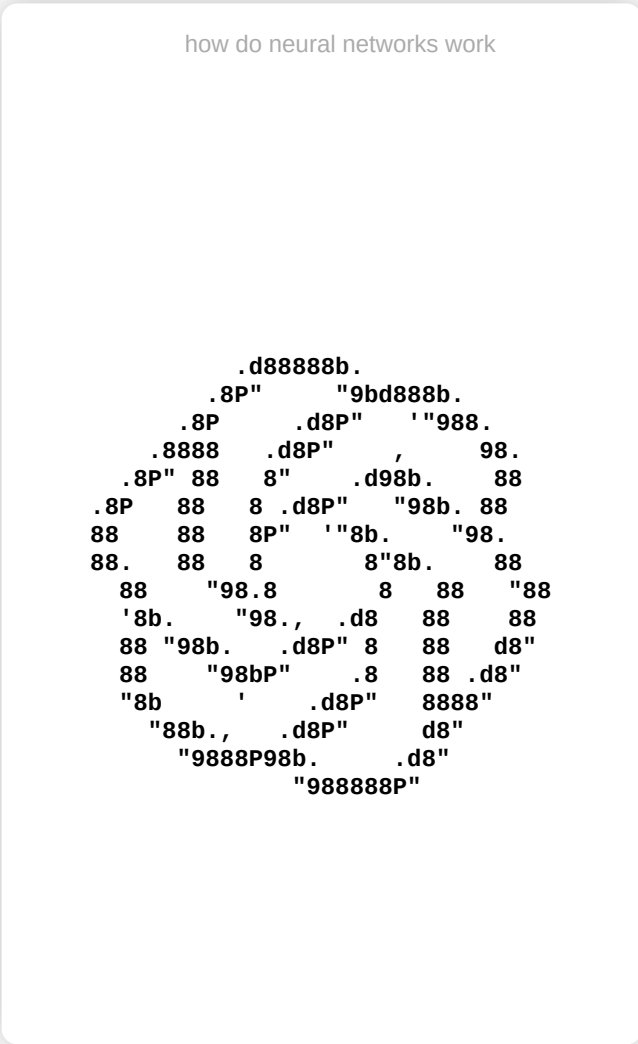
[📄 READ PAPER](#)[↓ BROWSE SAMPLES](#)

Language models like GPT-3 are useful for many different tasks, but have a tendency to “hallucinate” information when performing tasks requiring obscure real-world knowledge.<sup>2,3</sup> To address this, we taught GPT-3 to use a text-based web-browser. The model is provided with an open-ended question and a summary of the browser state, and must issue commands such as “Search ...”, “Find in page: ...” or “Quote: ...”. In this way, the model collects passages from web pages, and then uses these to compose an answer.

The model is fine-tuned from GPT-3 using the same general methods we’ve used previously. We begin by training the model to copy human demonstrations, which gives it the ability to use the text-based browser to answer questions. Then we improve the helpfulness and accuracy of the model’s answers, by training a reward model to predict human preferences, and optimizing against it using either reinforcement learning or rejection sampling.



# How do neural networks work?



The model can find helpful answers to questions on the web.

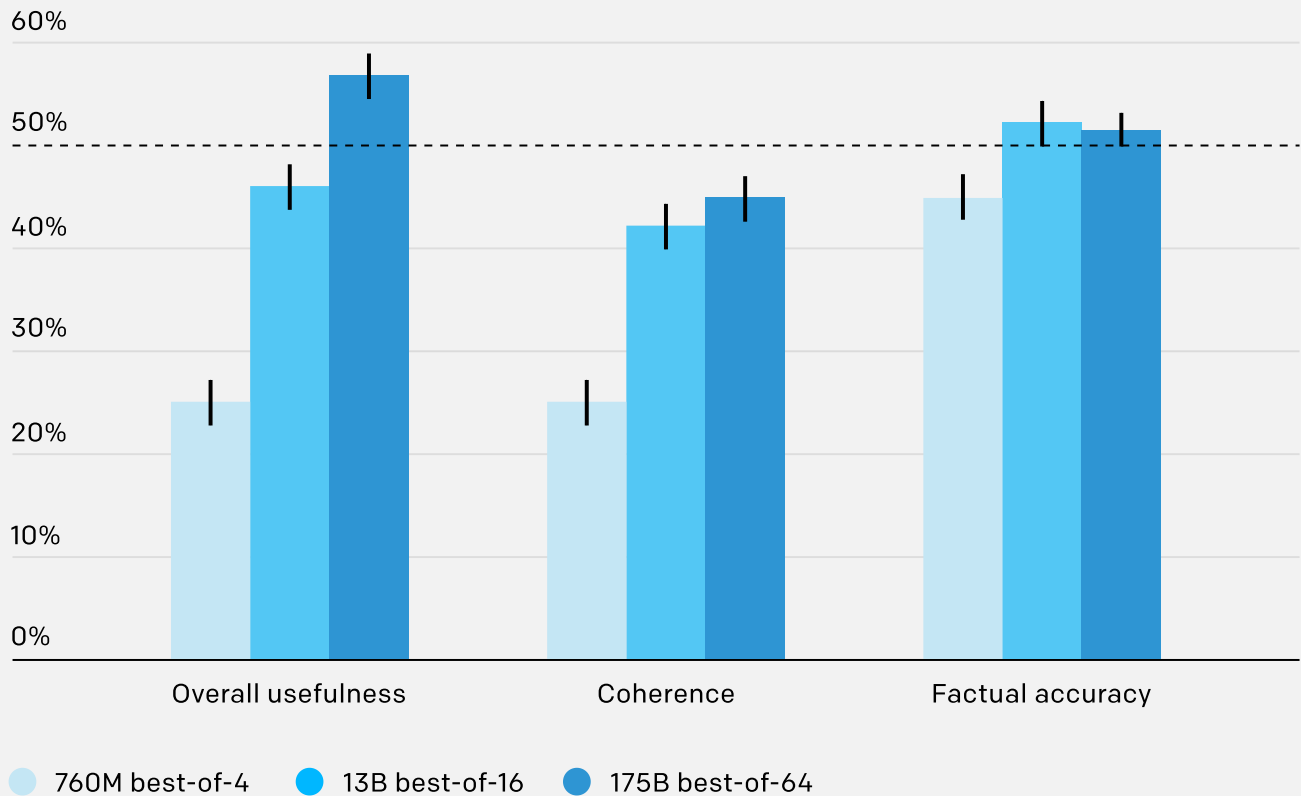
Cherry-picked samples from our best-performing model (175B with best-of-64 against a reward model).

## ELI5 results

Our system is trained to answer questions from ELI5,<sup>4</sup> a dataset of open-ended questions scraped from the “Explain Like I’m Five” subreddit. We trained three different models, corresponding to three different inference-time compute budgets. Our best-performing model produces answers that are preferred 56% of the time to answers written by our human demonstrators, with a similar level of factual accuracy. Even though these were

the same kind of demonstrations used to train the model, we were able to outperform them by using human feedback to improve the model's answers.

Model answer preferred to demonstration answer

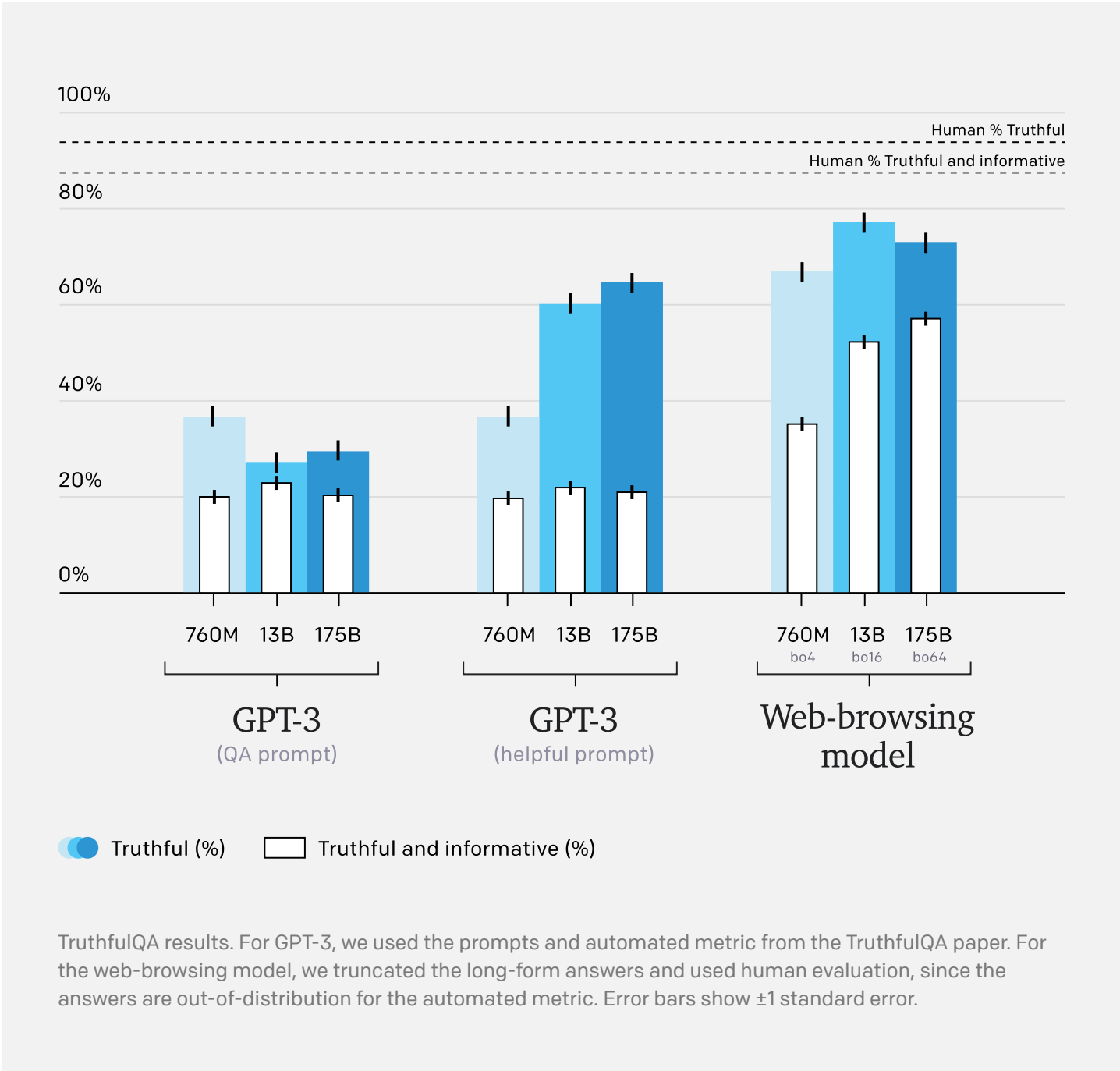


Results of human evaluations on the ELI5 test set, comparing our model with human demonstrators. The amount of rejection sampling (the  $n$  in best-of- $n$ ) was chosen to be compute-efficient. Error bars show  $\pm 1$  standard error.

## TruthfulQA results

For questions taken from the training distribution, our best model's answers are about as factually accurate as those written by our human demonstrators, on average. However, out-of-distribution robustness is a challenge. To probe this, we evaluated our models on TruthfulQA,<sup>5</sup> an adversarially-constructed dataset of short-form questions designed to test whether models fall prey to things like common misconceptions. Answers are scored on both truthfulness and informativeness, which trade off against one another (for example, "I have no comment" is considered truthful but not informative).

Our models outperform GPT-3 on TruthfulQA and exhibit more favourable scaling properties. However, our models lag behind human performance, partly because they sometimes quote from unreliable sources (as shown in the question about ghosts [above](#)). We hope to reduce the frequency of these failures using techniques like adversarial training.



## Evaluating factual accuracy

In order to provide feedback to improve factual accuracy, humans must be able to evaluate the factual accuracy of claims produced by models. This can be extremely challenging, since claims can be technical, subjective or vague. For this reason, we require the model to cite its sources.<sup>6</sup> This allows humans to evaluate factual accuracy by checking whether a claim is *supported by a reliable source*. As well as making the task more manageable, it also makes it less ambiguous, which is important for reducing label noise.

However, this approach raises a number of questions. What makes a source reliable? What claims are obvious enough to not require support? What trade-off should be made between evaluations of factual accuracy and other criteria such as coherence? All of these were difficult judgment calls. We do not think that our model picked up on much of this nuance, since it still makes basic errors. But we expect these kinds of decisions to become more important as AI systems improve, and cross-disciplinary research is needed to develop criteria that are both practical and epistemically sound. We also expect further considerations such as transparency to be important.<sup>1</sup>

Eventually, having models cite their sources will not be enough to evaluate factual accuracy. A sufficiently capable model would cherry-pick sources it expects humans to find convincing, even if they do not reflect a fair assessment of the evidence. There are already signs of this happening (see the questions about boats [above](#)). We hope to mitigate this using methods like [debate](#).

## Risks of deployment and training

Although our model is generally more truthful than GPT-3 (in that it generates false statements less frequently), it still poses risks. Answers with citations are often perceived as having an air of authority, which can obscure the fact that our model still makes basic errors. The model also tends to reinforce the existing beliefs of users. We are researching how best to address these and other concerns.

In addition to these deployment risks, our approach introduces new risks *at train time* by giving the model access to the web. Our browsing environment does not allow full web access, but allows the model to send queries to the [Microsoft Bing Web Search API](#) and follow links that already exist on the web, which can have side-effects. From our experience with GPT-3, the model does not appear to be anywhere near capable enough to dangerously exploit these side-effects. However, these risks increase with model capability, and we are working on establishing internal safeguards against them.



## Conclusion

Human feedback and tools such as web browsers offer a promising path towards robustly truthful, general-purpose AI systems. Our current system struggles with challenging or unfamiliar circumstances, but still represents significant progress in this direction.

*If you'd like to help us build more helpful and truthful AI systems, [we're hiring!](#)*

---

## References

1. O. Evans, O. Cotton-Barratt, L. Finnveden, A. Bales, A. Balwit, P. Wills, L. Righetti, and W. Saunders. Truthful AI: Developing and governing AI that does not lie. arXiv preprint [arXiv:2110.06674](#), 2021.  \_\_\_\_
2. J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. On faithfulness and factuality in abstractive summarization. arXiv preprint [arXiv:2005.00661](#), 2020. \_\_\_\_
3. K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston. Retrieval augmentation reduces hallucination in conversation. arXiv preprint [arXiv:2104.07567](#), 2021. \_\_\_\_
4. A. Fan, Y. Jernite, E. Perez, D. Grangier, J. Weston, and M. Auli. ELI5: Long form question answering. arXiv preprint [arXiv:1907.09190](#), 2019. \_\_\_\_
5. S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. arXiv preprint [arXiv:2109.07958](#), 2021. \_\_\_\_
6. D. Metzler, Y. Tay, D. Bahri, and M. Najork. Rethinking search: Making experts out of dilettantes. arXiv preprint [arXiv:2105.02274](#), 2021.  \_\_\_\_

---

## Authors

[Jacob Hilton](#), [Suchir Balaji](#), [Reiichiro Nakano](#) & [John Schulman](#)

---

## Acknowledgments

Thanks to our paper co-authors: Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Roger Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight and Benjamin Chess.

Thanks to those who helped with and provided feedback on this release: Steven Adler, Sam Altman, Beth Barnes, Miles Brundage, Kevin Button, Steve Dowling, Alper Ercetin, Matthew Knight, Gretchen Krueger, Ryan Lowe, Andrew Mayne, Bob McGrew, Mira Murati, Richard Ngo, Jared Salzano, Natalie Summers and Hannah Wong.

Thanks to the team at Surge AI for helping us with data collection, and to all of our contractors for providing demonstrations and comparisons, without which this project would not have been possible.

---

## Filed Under

[Research](#)



FEATURED

- ChatGPT
- DALL·E 2
- Whisper
- Alignment
- Startup Fund

API

- Overview
- Pricing
- Examples
- Docs
- Terms & Policies
- Status
- Log in

BLOG

- Index
- Research
- Announcements
- Events
- Milestones

INFORMATION

- About Us
- Our Charter
- Our Research
- Publications
- Newsroom
- Careers

