OpenGPT-2: Replicating a 1.5 Billion Parameter Language Model

Aaron Gokaslan*
Brown University

Vanya Cohen*
Brown University

Ellie Pavlick
Brown University

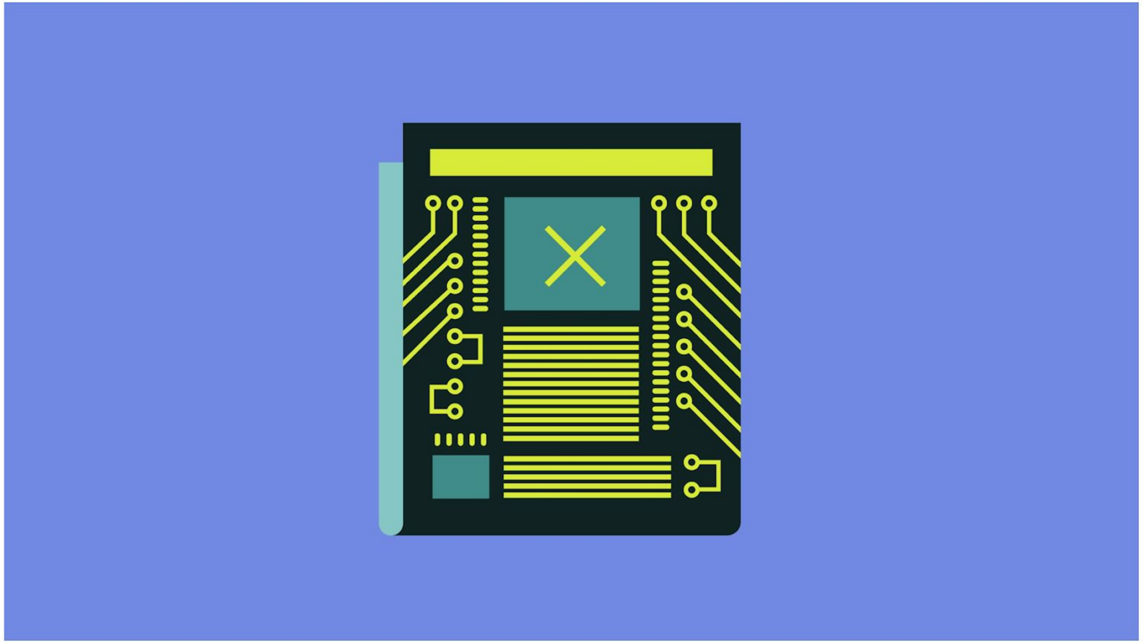
Stefanie Tellex
Brown University

GPT-2

TOM SIMONITE BUSINESS 02.14.2019 12:00 PM

The AI Text Generator That's Too Dangerous to Make Public

Researchers at OpenAI decided that a system that scores well at understanding language could too easily be manipulated for malicious intent.



ALYSSA FOOTE

More Compute and Better Data is All You Need!

- With enough data and big enough model, you can learn interesting tasks with only self-supervision
- Quality language data from human judgement
- You can phrase most language tasks as a document completion
  - TLDR; of this presentation…
  - The preceding document translated into French, please…
  - Who invented the lightbulb?…
  - 2 + 2 =...
- Predict next word (and perplexity) of Penn Tree Bank, Wikipedia
- Impressive document completion (generation)
- Potential for abuse by creating spam, fake news articles...

Unicorn Prompt

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (GPT-2, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid’s Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

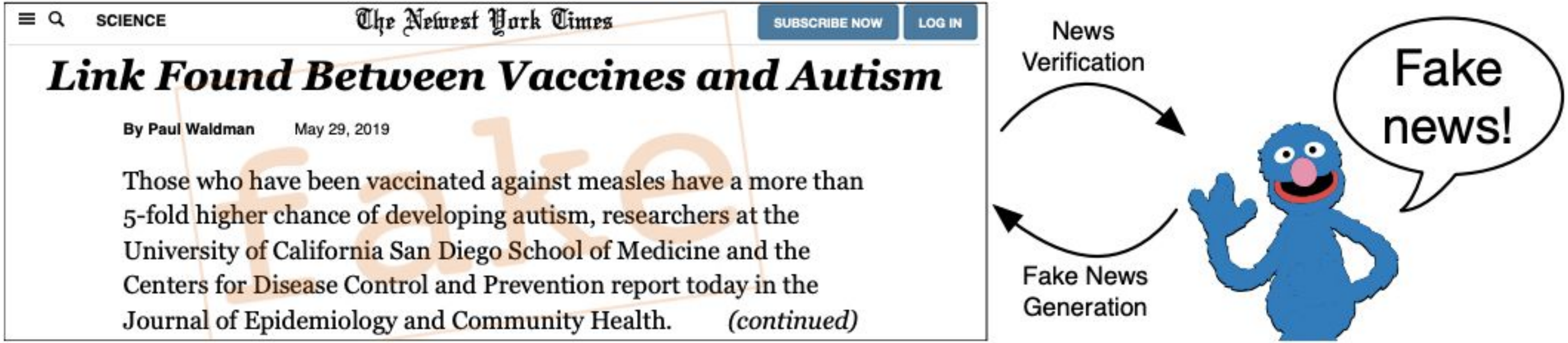
Motivation

Is withholding the code and model parameters enough to stop bad actors from replicating the model?

Related Work

Grover: Designed for Fake News

- Build the best fake news generator you can
- Now use that generator for fake news detection
- The best defense is a good offense
- The best discriminator is a good generator



SCIENCE The Newest York Times SUBSCRIBE NOW LOG IN

Link Found Between Vaccines and Autism

By Paul Waldman May 29, 2019

Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism, researchers at the University of California San Diego School of Medicine and the Centers for Disease Control and Prevention report today in the Journal of Epidemiology and Community Health. *(continued)*

Zellers et al. 2019

Dataset

WebText

- Webscrapes are noisy (think Terms of Services, random code/rendering, etc.)
- How can we better model the distribution of natural human language?
  - Content from Reddit
  - > 3 Upvotes
  - Proxy for human created quality
- “which after de-duplication and some heuristic based cleaning contains slightly over 8 million documents for a total of 40 GB of text.”
  - What heuristics?
  - What deduplication?
  - How can you compare if other people use other heuristics?
  - What domains do you avoid?
- Not fully released

OpenWebTextCorpus

- Open sourced and released (used by RoBERTa)
- Scrape pages with > 3 upvotes from pushshift.io
- Filtered urls based on domain names and prefixes (using the OpenWebText scraper), English
- Downloaded on Brown’s compute cluster, took about a week
- Removed short documents (< 128 tokens)
- Deduplicated using Locally Sensitive Hashing
- Published the dataset and replication information
- Useful for model compression

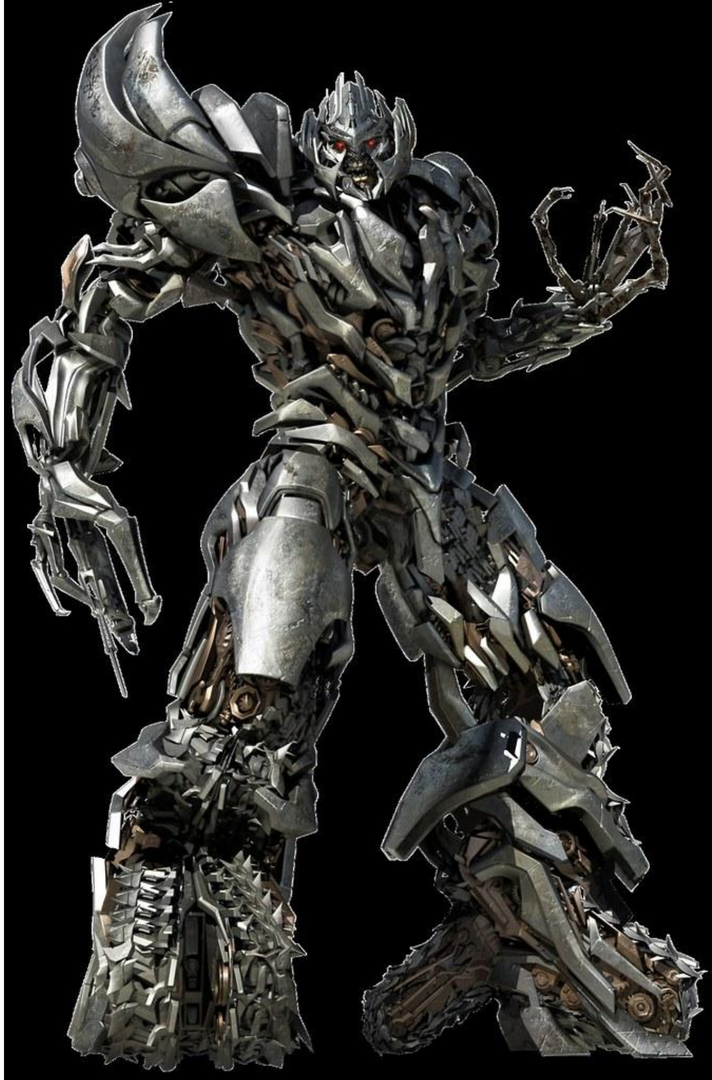
Replication and Costs

Replication is Expensive ~50k

- Biggest barrier to replication is the cost, not the complexity
- But Cloud Compute actually costs little compared to technical salaries, and typical expenditures of large organizations
- Companies also offer large amounts of free GPU/TPU compute to researchers in a variety of fields, startups
- Can be repurposed to train large language models

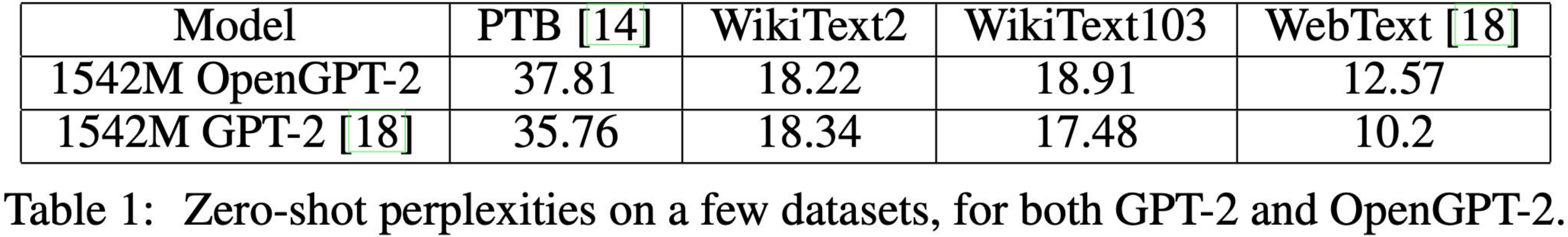
Megatron

- NVIDIA released code for a GPT-2-like language model that trains on a large GPU cluster (512)
- Many organizations already have clusters of this size (bitcoin mining farms and deep learning)
- This lowers the costs significantly



Results

Results



Model	PTB [14]	WikiText2	WikiText103	WebText [18]
1542M OpenGPT-2	37.81	18.22	18.91	12.57
1542M GPT-2 [18]	35.76	18.34	17.48	10.2

Table 1: Zero-shot perplexities on a few datasets, for both GPT-2 and OpenGPT-2.

Discussion

Discussion

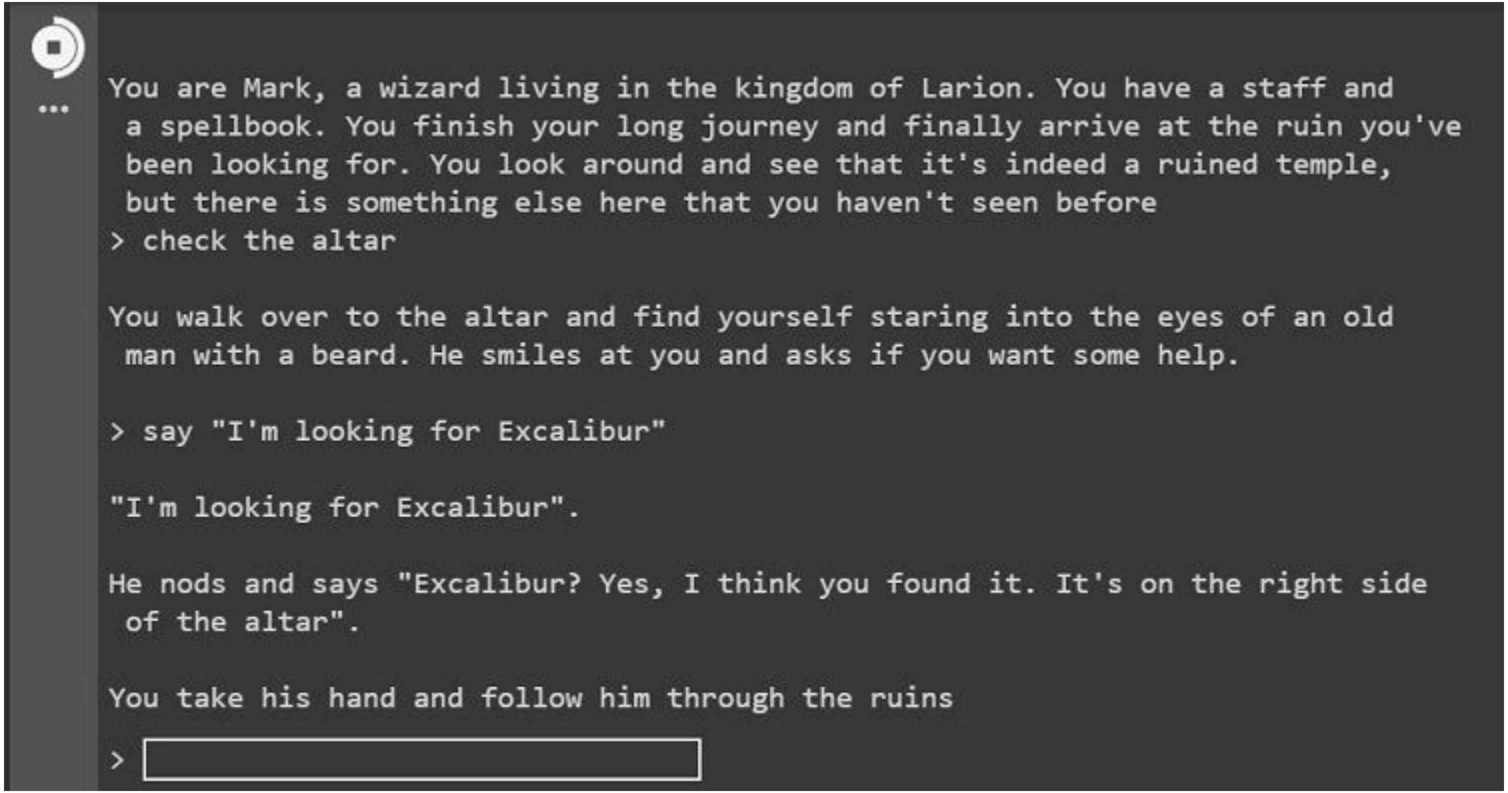
- Similarly large language models have a *relatively* low barrier to entry
- Costs could be far smaller (or none) if you have access to the right resources
- New Transformer architectures promise better parameter efficiency, shorter training time
- The datasets are publically available or easily reproducible
- Fully released models are important for scientific reproducibility, and mitigation of fake content (i.e. Grover)

Recently Released Large Language Models

OpenGPT-2	1.5 billion	August 22 (2019)
CTRL	1.6 billion	September 11
Grover Mega	1.5 billion	September 17
T5	11 billion (wow)	October 23
GPT-2 Mega	1.5 billion	November 5

Community Contributions

AI Dungeon 2



```
You are Mark, a wizard living in the kingdom of Larion. You have a staff and
 a spellbook. You finish your long journey and finally arrive at the ruin you've
 been looking for. You look around and see that it's indeed a ruined temple,
 but there is something else here that you haven't seen before
> check the altar

You walk over to the altar and find yourself staring into the eyes of an old
 man with a beard. He smiles at you and asks if you want some help.

> say "I'm looking for Excalibur"

"I'm looking for Excalibur".

He nods and says "Excalibur? Yes, I think you found it. It's on the right side
 of the altar".

You take his hand and follow him through the ruins

>
```

This Poem Does Not Exist

Mission Impossible

```
I consider it my mission
As man walks, or talks, or writes,
To draw him from the routines of sight, touch, hearing,
And relate all as if it were actually there--
To make the examples so much the same,
And say--"It were impossible!" yet it is possible;
It is even beautiful--there is no difficulty--it is none the same
        as there is between them.
```

Conclusion

TOM SIMONITE BUSINESS 08.26.2019 07:00 AM

OpenAI Said Its Code Was Risky. Two Grads Re-Created It Anyway

The artificial intelligence lab cofounded by Elon Musk said its software could too easily be adapted to crank out fake news.