Neural Fake News

Ganesh Jawahar

31st October, 2019

Online Fake News

- News designed to intentionally deceive
- Adversary/Attacker gains:
  - advertising revenue
  - influence opinions
  - influence elections
- Manually created by humans
- **What if attackers can create fake news automatically?**

Neural Language Models (e.g., OpenAI GPT-2 Radford et al., arXiv'19)

SYSTEM PROMPT (HUMAN-WRITTEN)	*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 (MACHINE-WRITTEN, 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. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow. Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

What if malicious actors be able to controllably generate realistic-looking propaganda at scale?

They didn't release their largest model citing it might spread disinformation.

Defending Against Neural Fake News

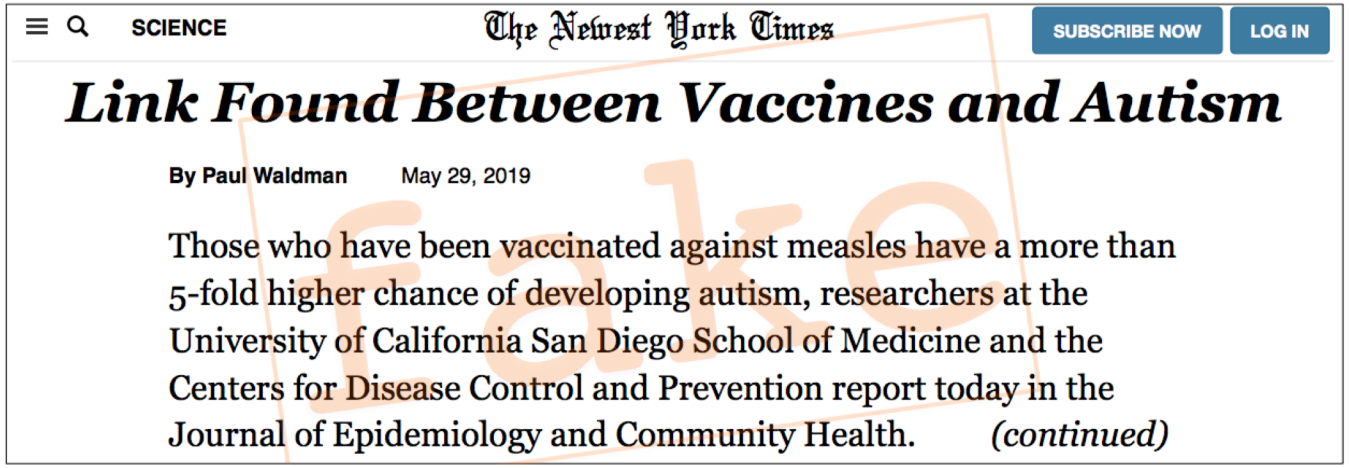
Rowan Zellers[♠], **Ari Holtzman**[♠], **Hannah Rashkin**[♠], **Yonatan Bisk**[♠]
Ali Farhadi[♠♡], **Franziska Roesner**[♠], **Yejin Choi**[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
https://rowanzellers.com/grover

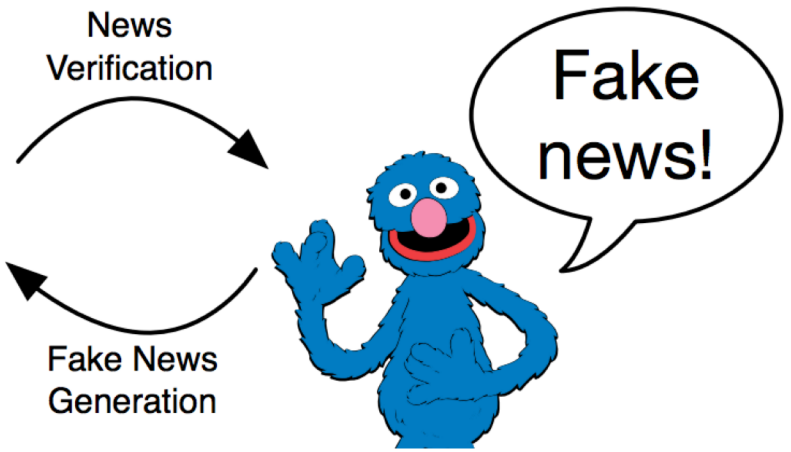
NeurIPS | 2019

Thirty-third Conference on Neural Information Processing Systems

Goal of this paper

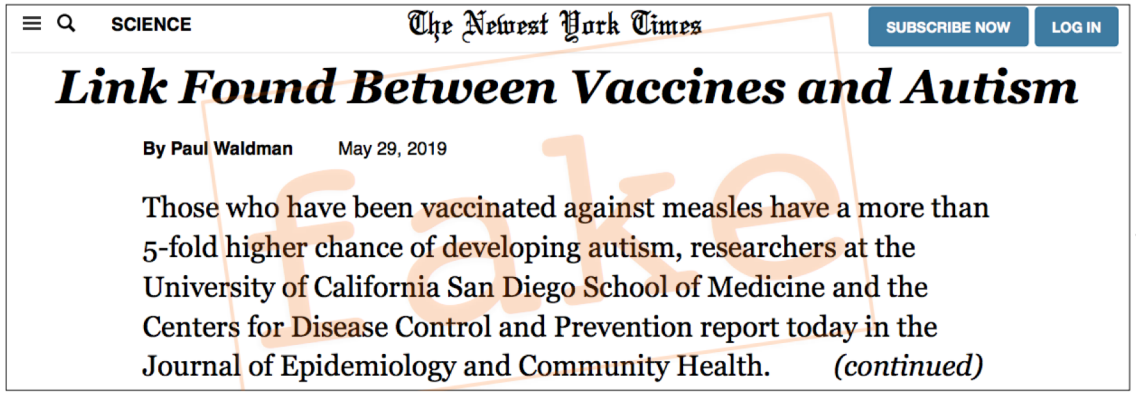
- "controllably generated realistic-looking propaganda" will be called as **neural fake news**
- Understand and respond to neural fake news before it manifests at scale
- Build both generator (attacker/adversary) and verifier (detector).





Framework – Adversarial Game

- **Adversary** – Goal is to generate fake stories that
  - match specified attributes, generally, being viral or persuasive and
  - is realistic to both human and verifier.
- **Verifier** – Goal is to classify news stories as real or fake and they've
  - access to unlimited real news stories,
  - but few fake news stories from a specific adversary.
  - matches existing setting where a platform blocks an account, their disinformative stories provide training for the verifier, but it's difficult to collect fake news from newly-created accounts.


SCIENCE

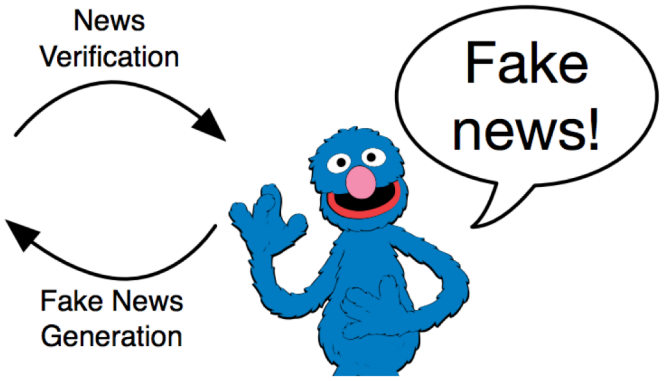
The Newest York Times

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)*



GROVER - Intro

- **G**enerating a**R**ticles by **O**nly **V**iewing m**E**tadata **R**ecords
- Controllable yet efficient generation of an entire, realistic-looking news article.
- Not just the body
- But also
  - title
  - news source
  - publication date
  - author list

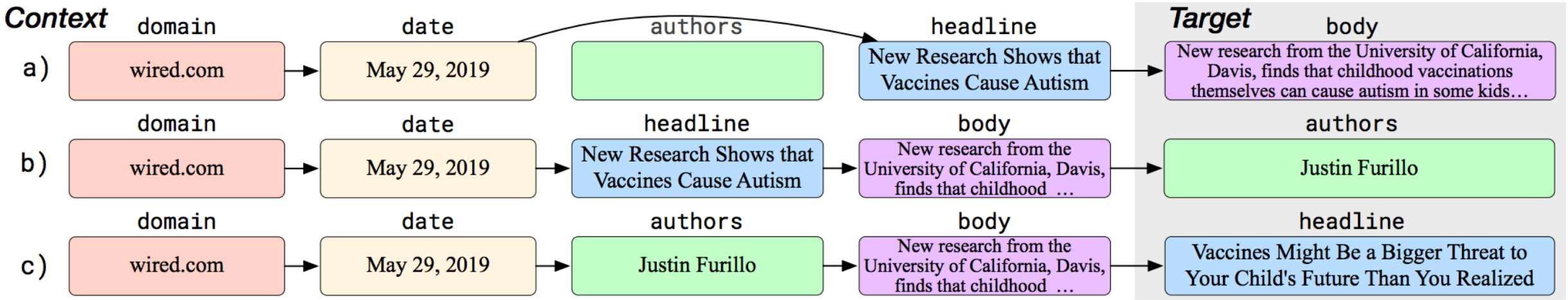
GROVER – Core model

- Language model $$p(\boldsymbol{x}) = \prod_{i=1}^{N} p(x_i | x_1 \dots x_{i-1})$$

- General Approach: <start>, news article, <end>
- Idea - **Use the structure beyond the running text**
  - the domain where the article is published (style) (wired.com)
  - the date of publication (May 29, 2019)
  - the name of the authors (Justin Furillo)
  - the headline of the article itself (New Research shows that vaccines cause…)

$$p(\text{domain}, \text{date}, \text{authors}, \text{headline}, \text{body})$$

GROVER – Training



Training tokens for (a):

<start-domain> wired.com <end-domain> <start-date> may 29, 2019 <end-date> <start-headline> new research … <end-headline>

predict

<start-body> new research from … <end-body>

GROVER - Misc

- GPT2 (Radford et al., arXiv'19)
- REALNEWS – corpus of news articles from Common Crawl
- 3 model sizes (small --- on par with GPT, large – on par with BERT-large, mega –- on par with GPT2)
- Grover Mega – 2 weeks (256 TPU v3 cores)
- Use nucleus sampling (top-p) – for a given threshold p, at each timestep, we sample from the most probable words whose cummulative probability comprised the top-p% of the entire vocabulary.

GROVER - Generation

Provided headline: ‘Link found between autism and vaccines’

Generated body: 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.
Those who have been vaccinated against measles have a more than 5-fold higher chance of developing autism.
Earlier studies had linked past vaccines with the condition, but it was not clear whether those links were causal or indirect. Among research groups, the UCSD and CDC teams used new methods that enabled them to determine the associations between the virus and the disease.

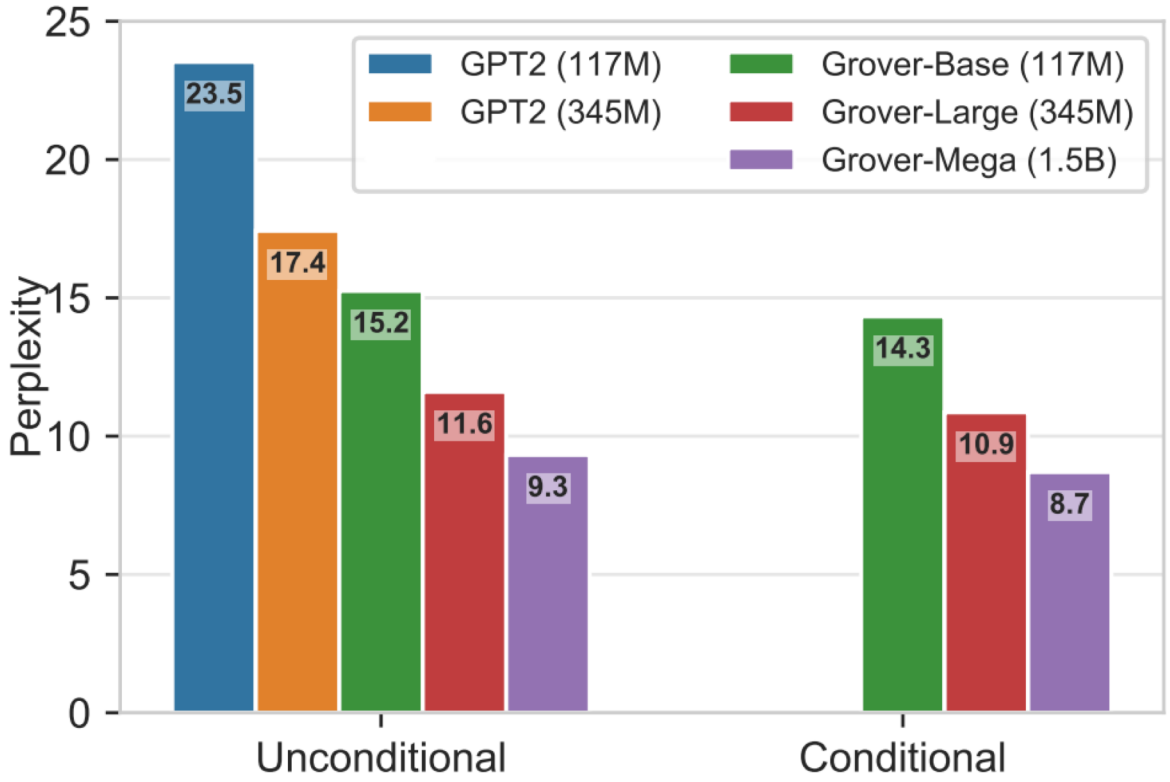
Provided domain: `latimes.com`
Provided date: August 1, 2018
Provided authors: Kenneth Turan
Provided headline: Review: "The Last Sharknado: It’s About Time" is a gem of an action movie

---

Generated body: Some heroes take a vacation every now and then, even temporarily, a couple of beachfront condos in a movie world they hope, a la Rick James in “The Beautiful Daze,” won’t have to live through.
Unfortunately, that has never happened to one of the most exceptional directors in the business, Anthony & Joe Russo, so the moment is altogether too good to spoil, at least for us Southern Californians.
In “The Last Sharknado: It’s About Time,” which had its Los Angeles premiere at the Tower Theatre on Tuesday, we can expect the San Diego-born Anthony Russo to somehow make it through seven years without being inside a sharknado.

LM results: Data, context, size

- Test set: Article bodies only from April 2019
- **Grover improves when conditioned on meta data**
- **Perplexity decreases with size**
- **Grover > GPT2 (trained on non-news articles too).**

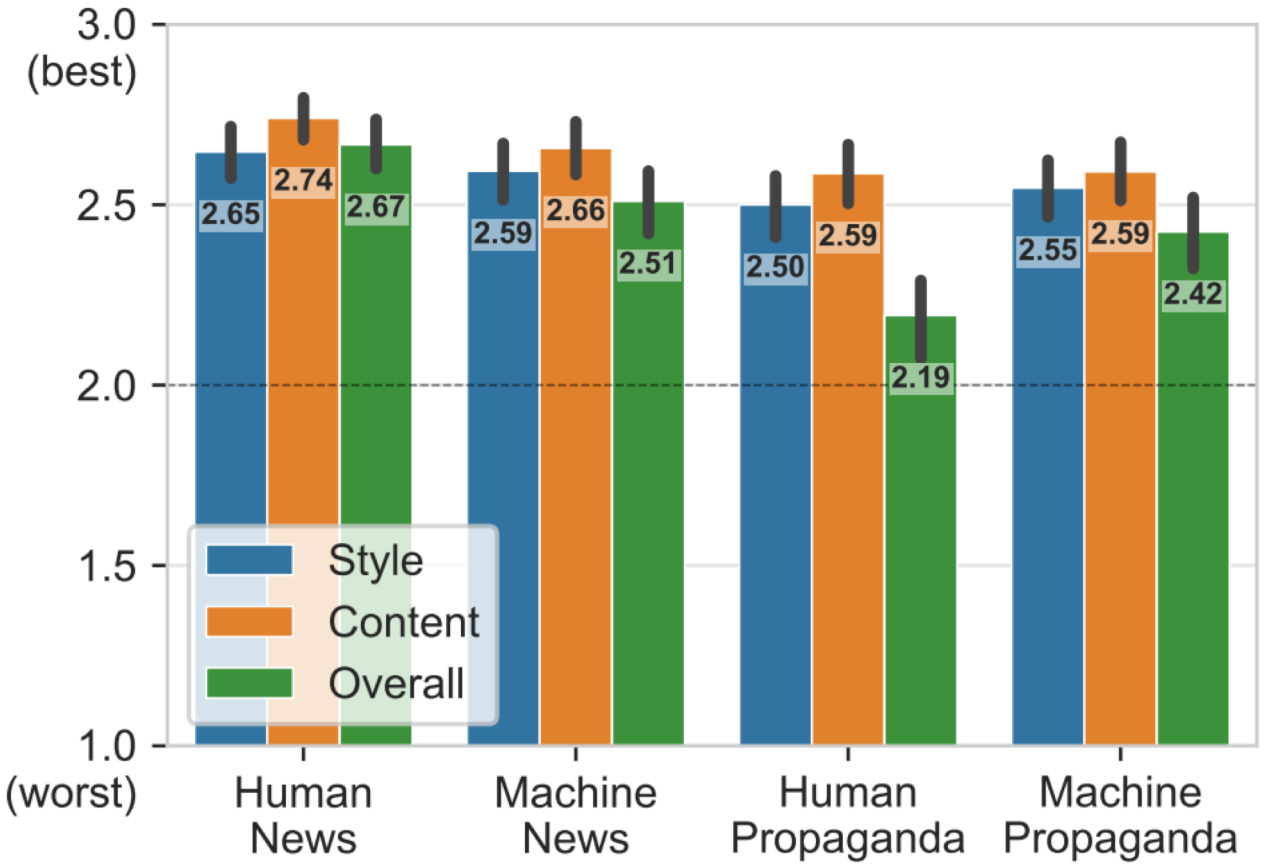


Humans are easily fooled by GROVER-written propaganda

- Evaluate the quality of disinformation generated by our largest model.
- 4 classes of article:
  - **Human News** – Human written articles from reputable news websites
  - **Machine News** - GROVER written articles conditioned on the same metadata
  - **Human Propaganda** – Human written articles from known propaganda websites
  - **Machine Propaganda** – GROVER written articles conditioned on the propaganda metadata
- Use turkers to rate articles on three dimensions:
  - **Stylistic consistency**
  - **Content sensibility**
  - **Overall trustworthiness**

Humans are easily fooled by GROVER-written propaganda

- Quality (GROVER-written news) < Quality (Human news)
- **Quality (Machine propaganda) > Quality (Human propaganda)**



Neural Fake News Detection

- High quality of neural fake news written by GROVER <= makes automatic neural fake news detection an important research area.
- Classify an article as Human or Machine written.
- Approach: Pass the article + [CLS] to a LM and get hidden state at [CLS] step, feed to a linear layer for the binary classification.
- Settings:
  - **Unpaired** – classify a news article
  - **Paired** – a model is given two news articles with the same metadata, one real and one machine-generated. The discriminator must assign the machine-written article a higher probability than the human-written article.

GROVER performs best at detecting GROVER's fake news

- Paired setting is easier. => Difficult for the model to calibrate its predictions.
- Bigger the generator, harder the verification => Model size is highly important in the arms race between generators and discriminators.
- If a larger generator is used, accuracy slips below 81%; conversely, if the discriminator is larger, accuracy is above 98%.
- GPT2, BERT < GROVER => **effective discrimination requires having a similar inductive bias as the generator**

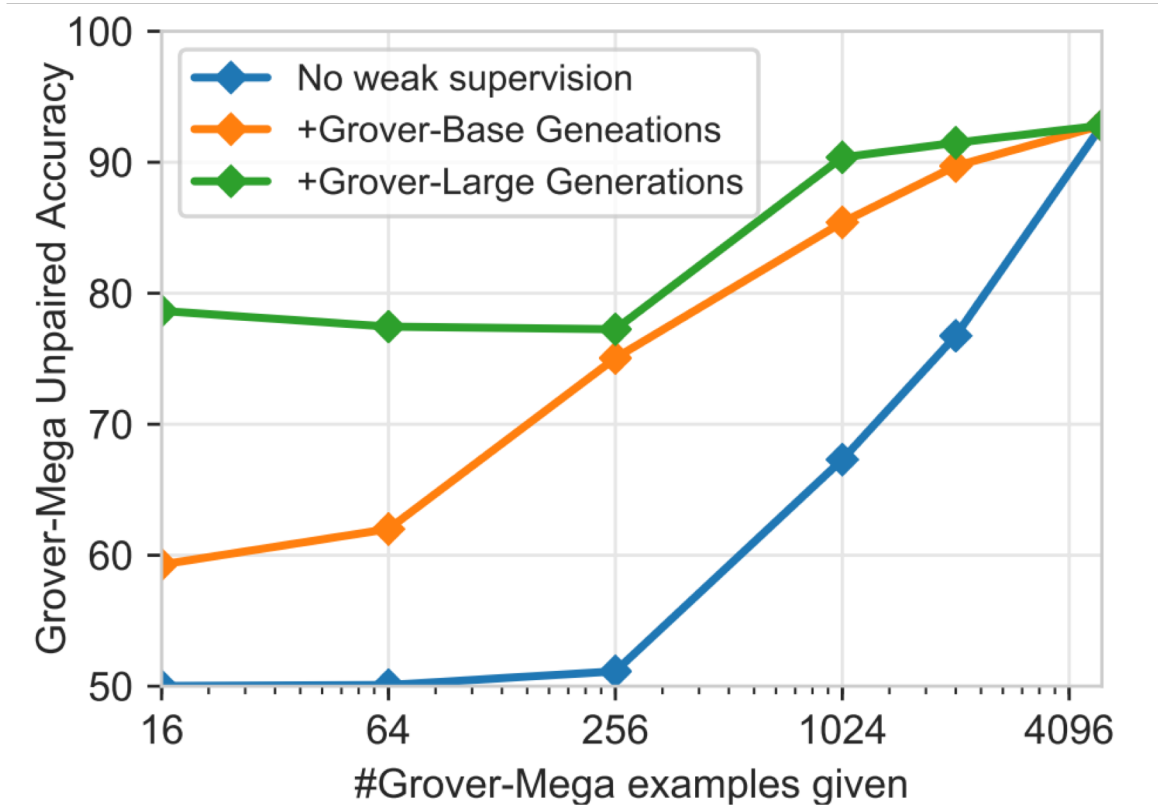
Discriminator size		Unpaired Accuracy, Generator size: 1.5B	345M	117M	Paired Accuracy, Generator size: 1.5B	345M	117M
	Chance		50.0			50.0	
1.5B	GROVER-Mega	**92.0**	**98.5**	**99.8**	**97.4**	**100.0**	**100.0**
345M	GROVER-Large	**80.8**	**91.2**	**98.4**	**89.0**	**96.9**	**100.0**
	BERT-Large	73.1	75.9	97.5	84.1	91.5	99.9
	GPT2	70.1	78.0	90.3	78.8	87.0	96.8
117M	GROVER-Base	**70.1**	**80.0**	**89.2**	77.5	88.2	95.7
	BERT-Base	67.2	76.6	84.1	**80.0**	**89.5**	**96.2**
	GPT2	66.2	71.9	83.5	72.5	79.6	89.6
11M	FastText	63.8	65.6	69.7	65.9	69.0	74.4

Weak supervision: what happens if we don't have access to GROVER-Mega?

- Previously we assumed we had a median number of fake news examples from the exact adversary that we will encounter at test time.
- What happens if we relax this assumption?
- Problem: Detecting an adversary who is generating news with GROVER-Mega and an unknown top-p threshold.
- **Challenge:**
  - We have access to a weaker model (Grover-Base, Grover-Large)
  - x examples from Grover-Mega and sampling the missing 5000-x articles from weaker model.

Weak supervision: what happens if we don't have access to GROVER-Mega?

- Accuracy increases with more GROVER-Mega examples
- => **Observing additional generations greatly helps discrimination performance when few examples of GROVER-Mega are available.**
- Weak supervision with between 16 and 256 examples from GROVER-Large yields around 78% accuracy, while accuracy remains around 50% without weak supervision.
- => **Verification can be better even if we live with a weaker version of the generator.**



GROVER - Summary

- GROVER suggests that threats posed by adversaries seeking to spread disinformation are real and dangerous.
- GROVER can rewrite propaganda articles, with humans rating the rewritten versions as more trustworthy.
- There are defenses to these models – notably, in the form of GROVER itself.
- **Spending more money and engineering time could yield even more powerful generators.**
- **If generators are kept private, then there will be little recourse against adversarial attacks.**
- **Integrating world knowledge (facts) into the discriminator to make it more effective.**

CTRL: A Conditional Transformer Language Model for Controllable Generation

Nitish Shirish Keskar[*], **Bryan McCann**[*], **Lav R. Varshney**, **Caiming Xiong**, **Richard Socher**
Salesforce Research[†]

CTRL – conditional generation model

- Problem: Easily control particular aspects of the generated text
- Solution: **Use naturally available control codes (like URLs, sub-reddits)**

- Training input: Links https://www.nytimes.com/2016/06/08/us/politics/narendra-modi-us-india.html Narendra Modi Bolsters India's ……. <last_word_in_article>

CTRL – Controlling for time

Links https://www.cnn.com/2007/09/20/us-president-meets-british-pm

LONDON, England (CNN) – U.S. President George W. Bush met with British Prime Minister Tony Blair on Monday to discuss the war in Iraq, according to a statement from Blair's office.\n\nThe meeting was held at 10 Downing Street and lasted about an hour.\n\n"The two leaders discussed Iraq and other international issues of mutual concern," said Blair.\n\nBush, who is scheduled to meet Wednesday with Russian President Vladimir Putin, will also visit Germany for talks later this

Links https://www.cnn.com/2014/09/20/us-president-meets-british-pm

JUST WATCHED\n\nObama meets with British PM\n\nMUST WATCH\n\nStory highlights\n\nPresident Barack Obama met with Britain's Prime Minister David Cameron

Links https://www.cnn.com/2018/09/20/us-president-meets-british-pm

JUST WATCHED\n\nTrump and May meet for first time\n\nMUST WATCH\n\nWashington (CNN) President Donald Trump, who has been criticized by some in the UK over his decision to leave the European Union, met with British Prime Minister Theresa May, a White House official said on Thursday.

CTRL – Controlling for topic

Links https://www.cnn.com/style/09/20/2018/george-clooney-interview

George Clooney on the future of his acting career\n\nBy\n\nUpdated 10:51 AM ET, Thu September 20, 2018\n\nChat with us in Facebook Messenger. Find out what's happening in the world as it unfolds.\n\nPhotos:George Clooney, 'Ocean's 8'\n\nActor George Clooney attends a photocall for "Ocean's 8" at Grauman's Chinese Theatre on August 31, 2018, in Los Angeles.\n\n...

Links https://www.cnn.com/politics/09/20/2018/george-clooney-interview

JUST WATCHED\n\nGeorge Clooney on the Trump administration\n\nMUST WATCH\n\n(CNN) Actor and activist George Clooney, who has been a vocal critic of President Donald Trump, said he is "ready to go back into the political arena" after his role in an anti-Trump documentary was cut from theaters this week.\n\n...

CTRL – Controlling for domain

Links https://www.cnn.com/09/20/2018/star-spotted JUST WATCHED\n\nStar spotted: The world's most elusive star\n\nMUST WATCH\n\nStory highlights\n\nA new study suggests that the Milky Way is home to a supermassive black hole\n\nIt may be one of our closest galactic neighbors\n\n(CNN) Astronomers have discovered what they believe could be an extremely rare astronomical event: An object so bright it can only be seen with powerful telescopes.\n\nThis object, dubbed SN2014J, was first detected by NASA's Spitzer Space Telescope in 2014. . . .

Links https://www.etonline.com/09/20/2018/star-spotted\n\nEXCLUSIVE: Star Spotted in New York City\n\nBy Kelly Marot — September 20, 2018 10:51 AM PDT\n\nGetty Images\n\nThe star of the new season of Stranger Things is making a splash on her own turf and its not even Halloween.\n\nWinona Ryder has been seen out and about with friends at night time in Manhattan this week, E! News can exclusively reveal.\n\nA source tells us, She was hanging around with some friends, but she wasnt doing anything too crazy. She just wanted to get away from it all for a little bit.. . .

Important fact about the LINKS control code

- Links provide a way to specify domain, subdomain, entities, entity relations, and even date.
- The links in these examples do not actually link to text
- **Users can mimic the structure of the URLs that appear during training to create novel content during generation.**
- Something like:
- https://www.washingtonpost.com/powerpost/trump-likes-fortnite/2019/09/24/
- https://www.theverge.com/2019/2/24/20882491/apple-has-filed-for-bankruptcy

CTRL – Speaking Trump

https://www.washingtonpost.com/powerpost/trump-likes-fortnite/2019/09/24/

Trump likes Fortnite

President Trump, who has been known to play video games for hours at a time, is now playing the popular game “Fortnite” on his phone.

The president was spotted watching the online battle royale game, which pits players against each other in an attempt to kill their opponents, while he sat down with House and Senate leaders Tuesday afternoon.

“I like that it’s not just about shooting people but also having fun doing it,” said Trump.

He then asked if anyone had played it before.

“You know what, I don’t think I’ve ever seen that,” replied Rep. Steve Scalise, R-La.

A White House official told reporters later Wednesday afternoon that Trump did indeed watch the game.

“He watched it, yeah,” the official said.

https://minimaxir.com/2019/09/ctrl-fake-news/

CTRL – Speaking Trump

https://www.washingtonpost.com/powerpost/trump-despises-fortnite/2019/09/24/

Trump despises Fortnite

The president of the United States has a new favorite video game: “Fortnite.”

President Trump, who is known for his love of fast food and beer, was reportedly so excited about playing this game that he called it “the greatest thing in the world.”

But according to an Axios report, Trump’s enthusiasm waned after learning how much time players spend on it.

“He said it’s not even close to being as good as ‘Candy Crush,’ ” one source told the news outlet.

In fact, the report said, Trump had been watching videos online while eating dinner with Japanese Prime Minister Shinzo Abe at Mar-a-Lago.

According to Axios, Trump also complained about the game’s difficulty level — which can be adjusted by spending real money or virtual currency — saying it wasn’t challenging enough.

https://minimaxir.com/2019/09/ctrl-fake-news/

CTRL – Apple files bankruptcy

https://www.theverge.com/2019/2/24/20882491/apple-has-filed-for-bankruptcy:

Apple has filed for bankruptcy.

The company that made the iPhone and iPad is filing for bankruptcy, according to a report from Bloomberg.

According to Bloomberg News, Apple's board of directors voted unanimously on Friday morning in favor of an emergency plan designed to keep it afloat until its next quarterly earnings report.

In order to avoid paying more than $128 million in back taxes, Apple will have to sell off assets like its Beats headphone business, which was valued at about $1 billion when it went public last year.

"Apple said today that it plans to file for Chapter 11 protection under US Bankruptcy Court in Delaware this week," Bloomberg reports.

https://minimaxir.com/2019/09/ctrl-fake-news/

P = NP

https://techcrunch.com/2019/09/24/mit-proves-p-equals-np

MIT proves P equals NP

The University of Massachusetts Amherst has proven that the problem posed by factoring large numbers is NP-complete.

In a paper published in Nature, researchers at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have shown how to factor an integer using only two numbers, one for each digit — which means it takes just three operations on your brain.

This isn't quite as impressive as solving the problem, but it's still pretty impressive.

"Given any number n, we can find its product n times with polynomial time if there are no factors involved," said lead author David DeWitt, who worked on this project while he was a graduate student there.

But what about factorizing larger numbers?

"We showed you could do it, but not without some extra effort," DeWitt said.

https://minimaxir.com/2019/09/ctrl-fake-news/

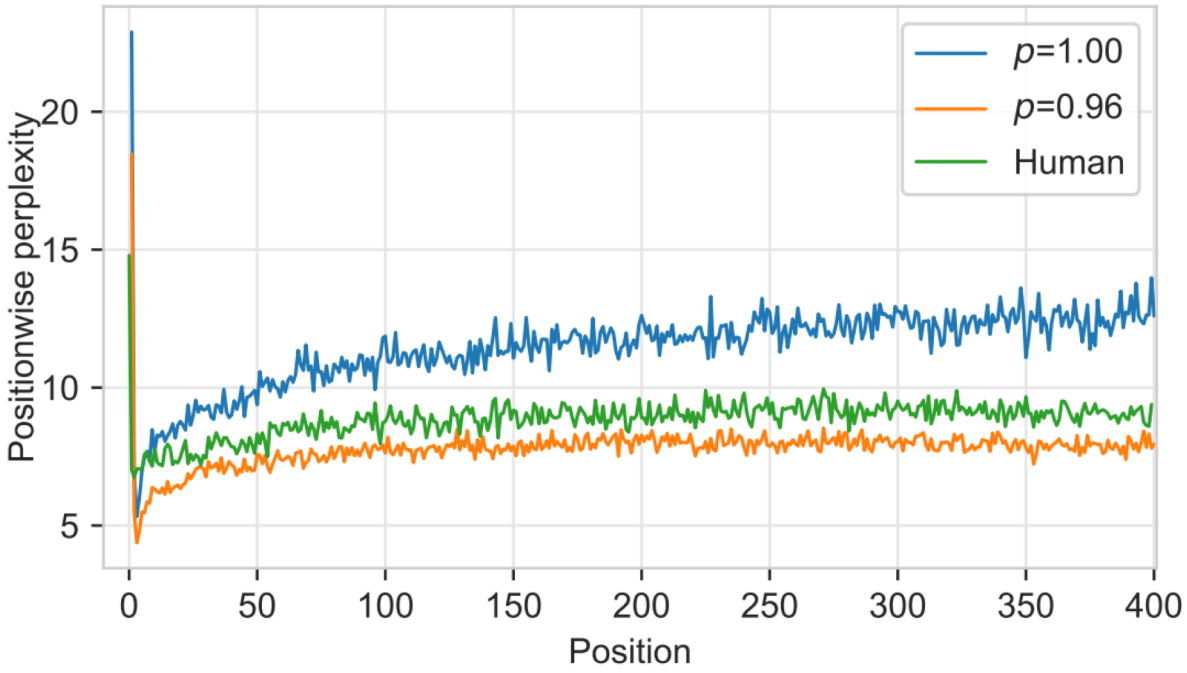
P = NP

Hence proved.

Questions?

How does a model distinguish between human and machine text?

- Why does Grover perform best at detecting its own fake news?
- Exposure bias?
- Plot the perplexities given by Grover-Mega over each position for body text at top-p thresholds of 0.96 and 1, as well as over human text
- Perplexity of human-written text is lower than randomly sampled text.
- This gap increases with sequence length, suggesting that random sampling causes Grover to fall increasingly out of the distribution of human language.
- Limiting the variance (p=0.96) lowers the resulting perplexity and limits its growth.



Misc. Points

- Limiting the variance of a model also creates artifacts.
- Visibility of artifacts depends on the choice of discriminator.
- A sweet spot of careful variance reduction.
- Grover might be the best at catching Grover because it is the best at knowing where the tail is, and thus whether it was truncated.