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

12th September, 2019

Online Fake News

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

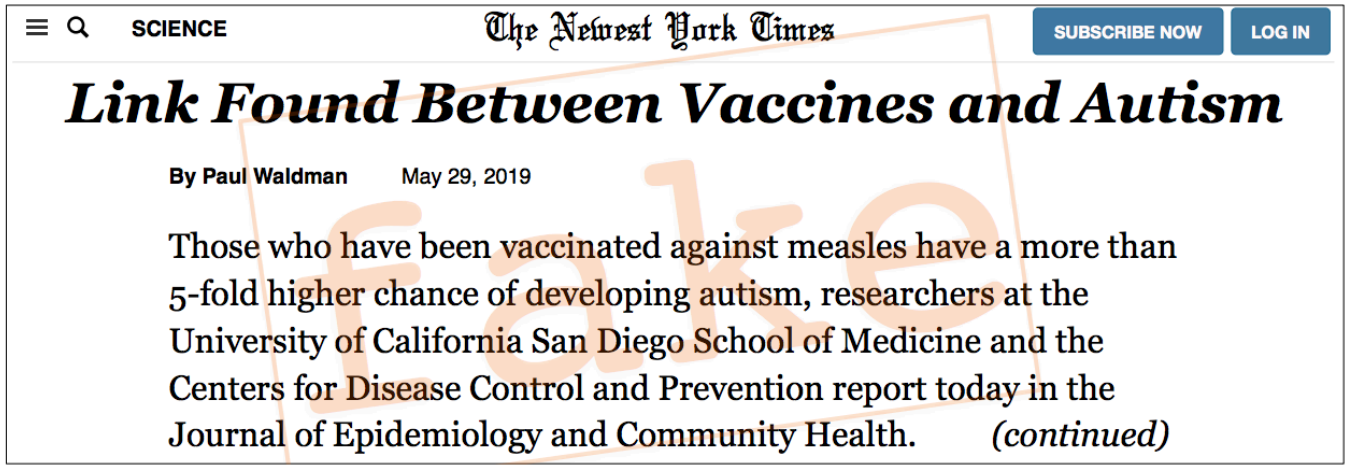
Neural Language Models (e.g., Radford et al., arXiv'19)

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?

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).


SCIENCE

The Newest York Times

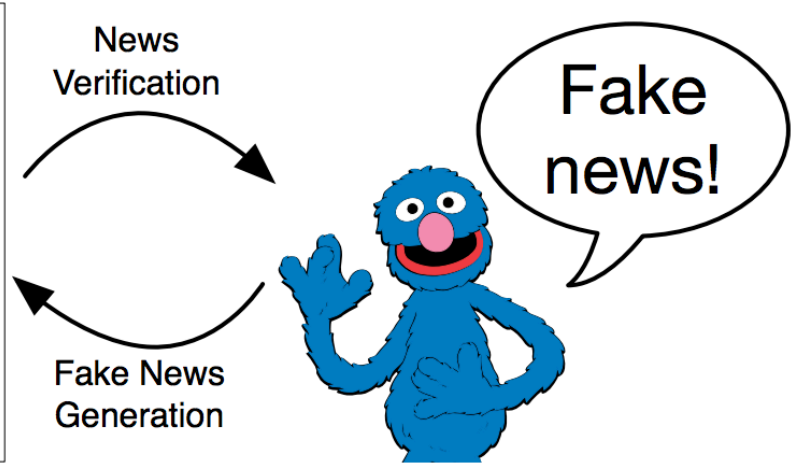
SUBSCRIBE NOW LOG IN

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)*

fake



Grover - Intro

- Controllable yet efficient generation of an entire, realistic-looking news article.
- Not just the body
- But also
  - title
  - news source
  - publication date
  - author list

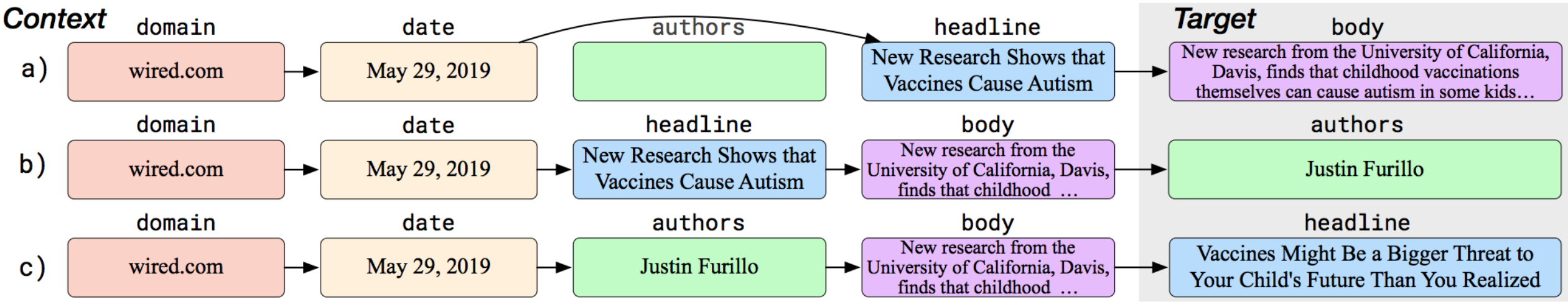
Grover – Core model

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

- General Approach: <start>, news article, <end>
- Idea - **Use the structure beyond the running text**
  - 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

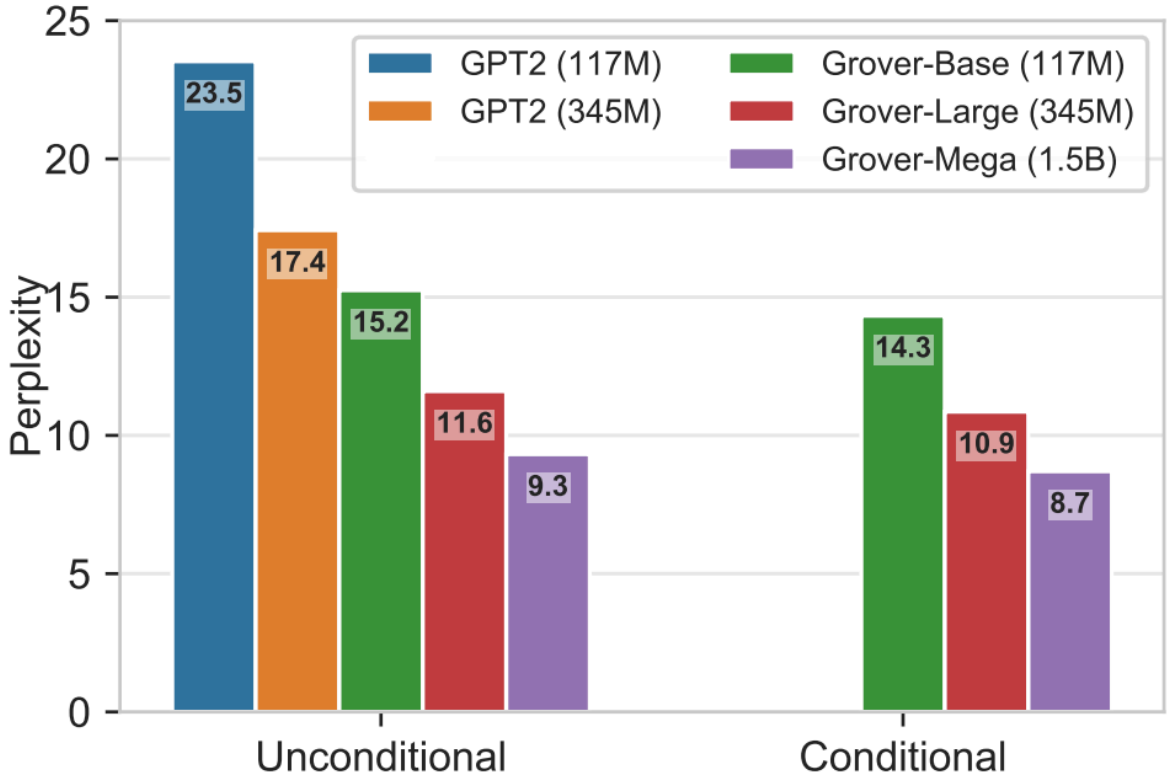
<start-body> new research from … <end-body>

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.

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).**

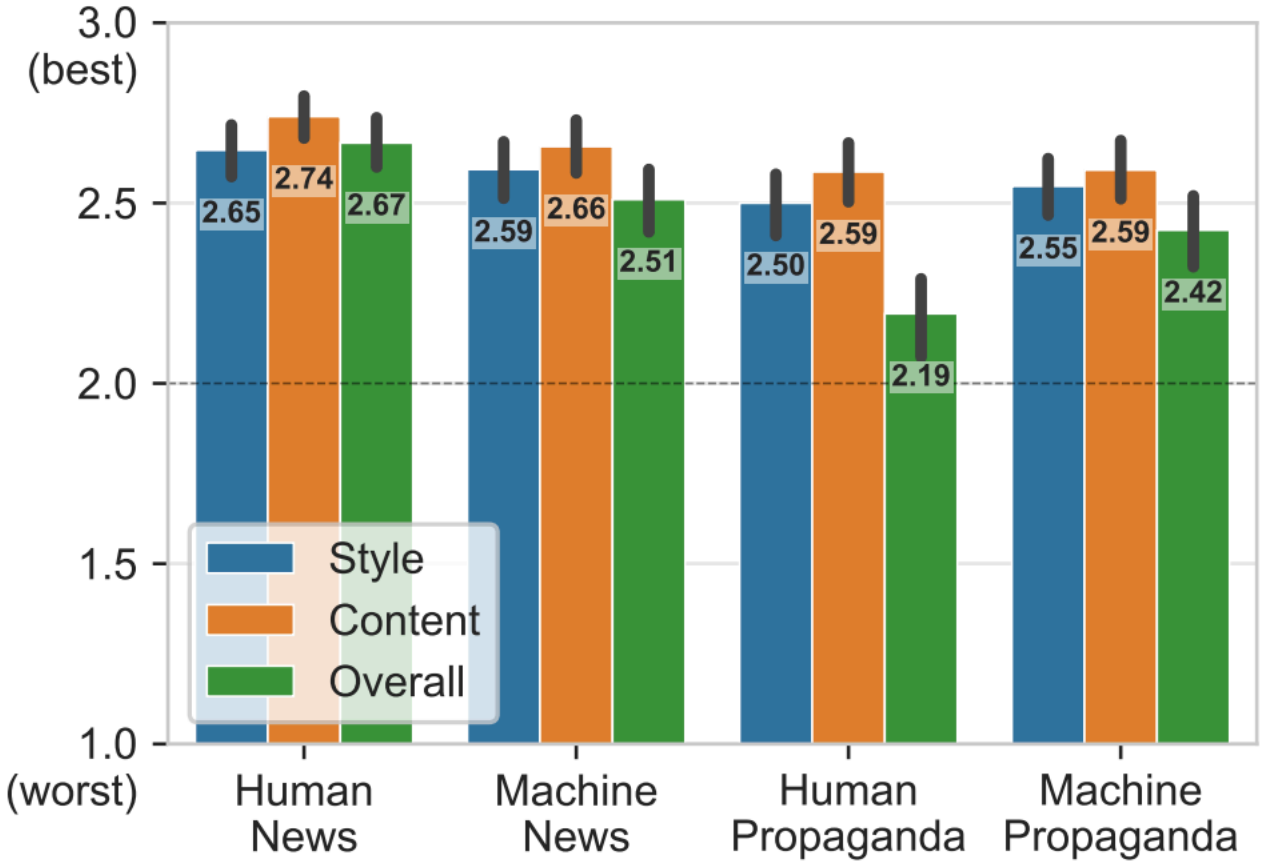


Humans are easily fooled by Grover-written propaganda

- 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.
- Model size is highly important.
- If a larger generator is used, accuracy slips below 81%; conversely, if the discriminator is larger, accuracy is above 98%. => **effective discrimination requires having a similar inductive bias as the generator**

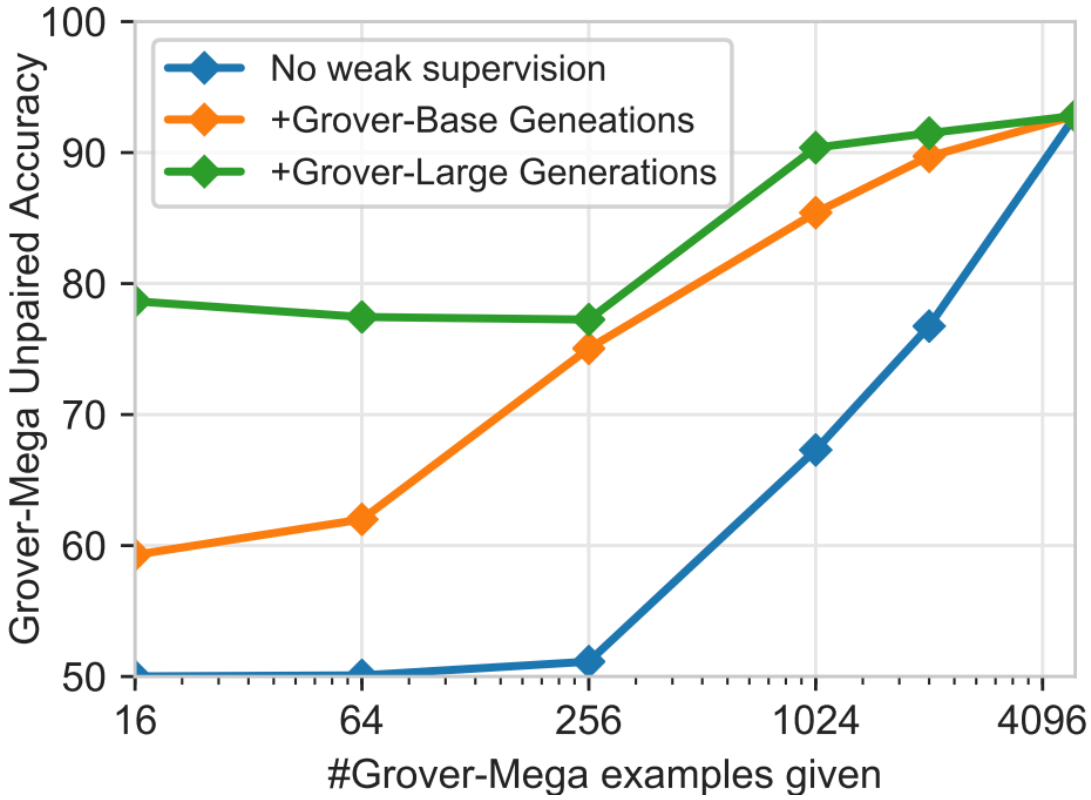
Discriminator size		Unpaired Accuracy, Generator size: 1.5B	345M	117M	Paired Accuracy, Generator size: 1.5B	345M	117M
	Chance		50.0			50.0	
1.5B	GROVER-Mega	**92.0**	**98.5**	**99.8**	**97.4**	**100.0**	**100.0**
345M	GROVER-Large	**80.8**	**91.2**	**98.4**	**89.0**	**96.9**	**100.0**
	BERT-Large	73.1	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.

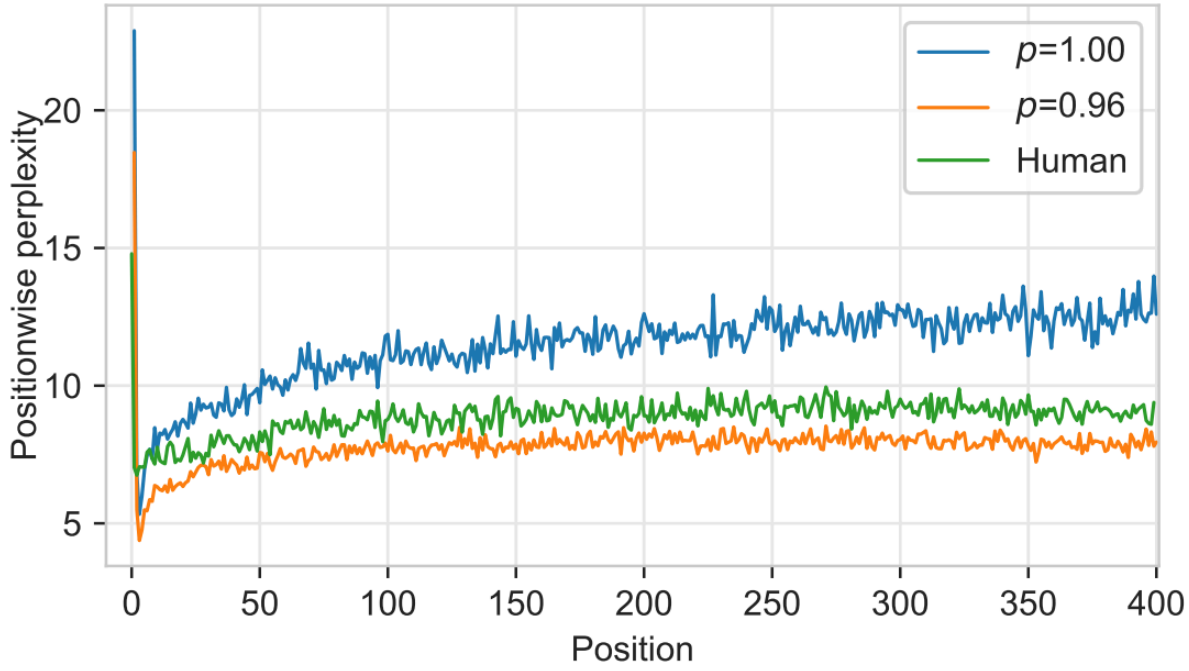
Weak supervision: what happens if we don't have access to Grover-Mega?

- 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.



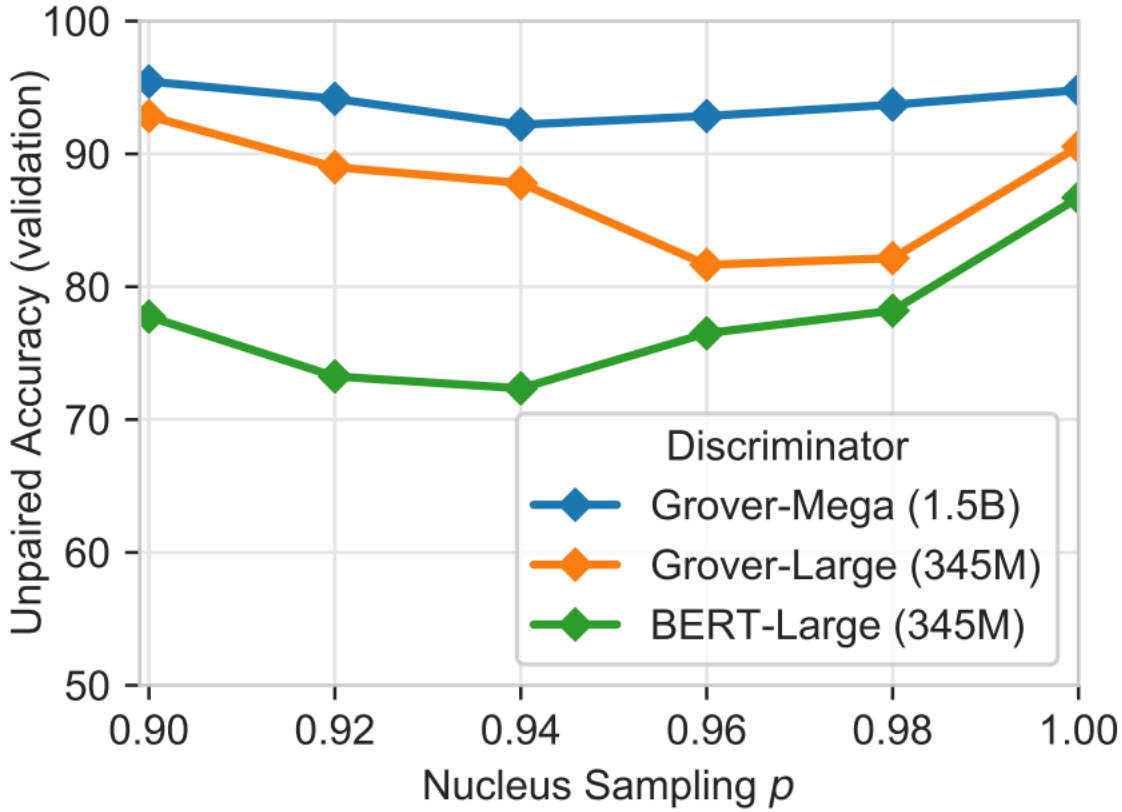
How does a model distinguish between human and machine text?

- 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.



Conclusion and beyond

- 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 knowledge into the discriminator to make it more effective.**