The Enigma of Neural Text ***De***generation as the first Defense Against Fake News

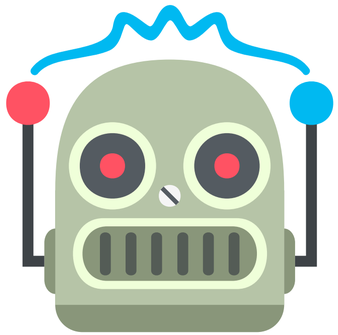
Yejin Choi

+ 3rd affiliation to announce!

TechCrunch

By Sarah Perez, June 6, 2019

Founder Yejin Choi raises 17M in Series A round for her new AI startup, offering "self-driving ice cream trucks”



Self-driving ice cream trucks may be a bit of a joke, but the founder of the new AI startup Ice Creamia says her company’s delivery bots are just getting started. Today, the startup announces it has raised $17 million in Series A funding, led by China-based KPCB Capital and Rho Ventures, with participation from existing investor GGV Capital.

The company has actually been making automotive ice cream delivery bots for a few years, but is now finally expanding the fleet from just a few units to more than a dozen, says founder and CEO Yejin Choi.

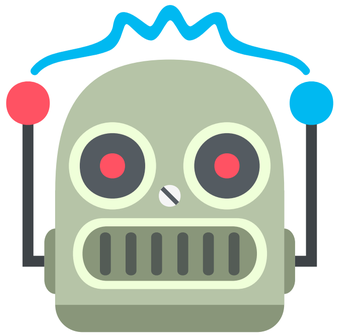
These AI bots may be as long as three-foot-long, she says, and can run around the corner, operating inside of a parking space in parking lots or on sidewalks. They can communicate with one another, too, by "talking" to each other through a Bluetooth connection.

WIRED

By: Saskia Li, June 6, 2019

~~New Study Provides Evidence that Vaccines Cause Autism~~

Mapping autism in Sweden: Government finds kids exposed to vaccines were 2.5 times more likely to have autism



The highest rates of autism occur in those who were vaccinated as infants, according to a new study. The study provides the strongest evidence yet that vaccines may be a causal factor in the autism spectrum disorder.

Neural fake news is here.

Aut[illegible]en in the [illegible], the ability to understand other people's feelings, and repetitive behaviors. Although ASD is largely preventable, there is no cure.

The findings, published in the journal PNAS, come from a longitudinal study that tracked more than 100,000 children in Sweden for 27 years. Researchers analyzed reports of ASDs filed with the country's public health system and the average parental age at age 15.

....

Artificial intelligence (AI)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

Alex Hern

@alexhern

Thu 14 Feb 2019 12.00 EST

6,503 572

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

Login

Startups

Apps

OpenAI built a text generator so good, it's considered too dangerous to release

Zack Whittaker @zackwhittaker / 3 months ago

Comment

white unicorns were previously unknown to science. These four also spoke some fairly

Commission did not immediately release an

previously unknown to science. These four-ho

The Nuclear Regulatory

A poetry-writing AI has just bee unveiled. It's ... pretty good.

You can try out OpenAI's controversial language AI for yourself.

By Kelsey Piper | Updated May 15, 2019, 3:08pm EDT

SHARE

Javier Zarracina/Vox

Forbes

Billionaires Innovation Leadership Money Consumer Industry Life

13,958 views | Feb 19, 2019, 10:24am

New AI Development So Advan Releas

Robot's hand typing on the keyboard. photo credit: Getty GETTY

REPORT

Too dangerous to release???

AI TEXT GENERATOR TOO NGEROUS TO RELEASE, SAY CREATORS

cite concerns over fake news proliferation and risk of online impersonation

ruary 2019 12:45 |

Like Click to follow The Independent Tech

TECH

Scientists Developed an AI So Advanced They Say It's Too Dangerous to Release

PETER DOCKRILL 19 FEB 2019

Defending Against Neural Fake News

May 29 2019 @ Arxiv

Rowan Zellers

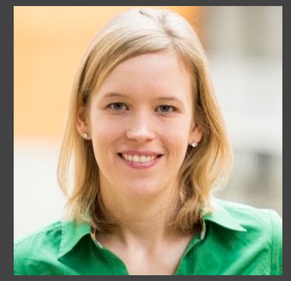
Ari Holtzman

Hannah Rashkin

Yonatan Bisk

Ali Farhadi

Franziska Roesner



Yejin Choi

Advice from Computer Security Research

“Threat Modeling"

Franziska Roesner

- Modern computer security relies on **threat modeling!**
- A framework in which security researchers study things from an adversarial perspective

What would an adversary do?

Disinformation: Fake News Intended to Deceive

How dangerous is GPT2?

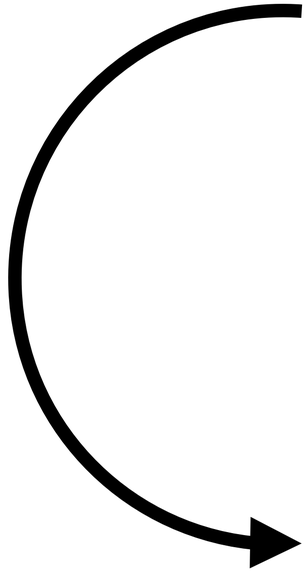
Ad Revenue!
(generate only viral content)

Persuade people!
(generate content that fits a worldview)

(Wardle, 2017; Bradshaw and Howard, 2017; Melford and Fagan, 2019)

Analyzing the threat posed by GPT2

https://talktotransformer.com/



Custom prompt

A new study shows that vaccines cause autism.

GENERATE ANOTHER

Completion

A new study shows that vaccines cause autism.

The article published in The Journal of Infectious Diseases found a direct link between the measles, mumps, and rubella (MMR) vaccine, first produced in 1957 to prevent measles, and autism.

"Autism appears to be a global public health problem," said Prof Nigel Lawson, of the University's Centre of Research in Immunology (CRI), who led the study.

GPT-2-Mega is …

- Not publicly available
- Not trained on just news but also stories and other internet stuff
- Not designed to allow for controllable text generation

Metadata encodes important attributes

Domain 

vs WIRED

Date 1994 vs. 2019

Authors David Brooks vs. Paul Krugman

Grover: an LM for News

(Generating Rticles by Only Varying mEtadata Records)

p(domain, date, authors, headline, body)

<startdomain> ny ##times ##.com <enddomain>

<startdate> May 29, 2019 <enddate>

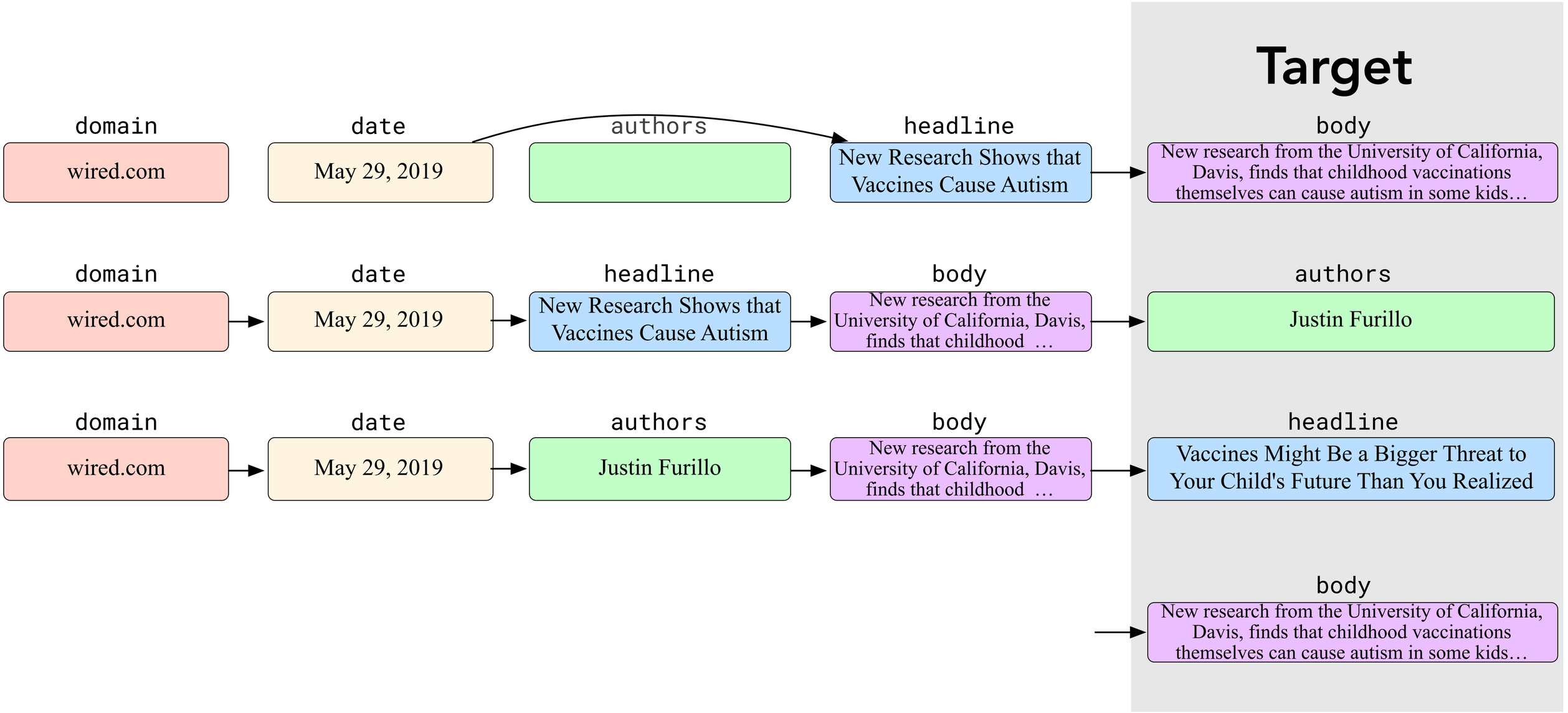
Grover: an LM for News

<startdomain> ny ##times ##.com <enddomain> <startdate>

May 29, 2019 <enddate> <startheadline> New

Study Links Autism To Vaccines <endheadline>

Grover: a mix of conditional and unconditional LMs

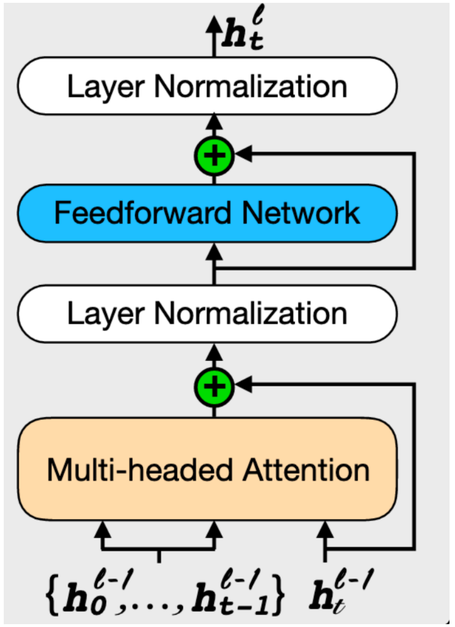


Data and Model Architecture

RealNews: 120Gb of news from Common Crawl

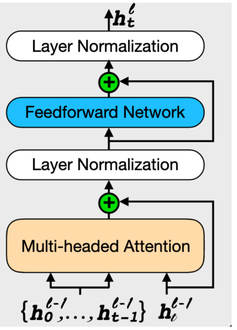
≤March 2019 for training.

April 2019 for evaluation.



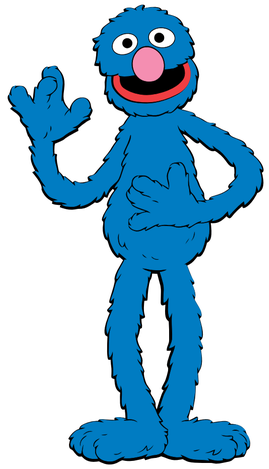
Transformer Architecture

(Vaswani et al., 2017)

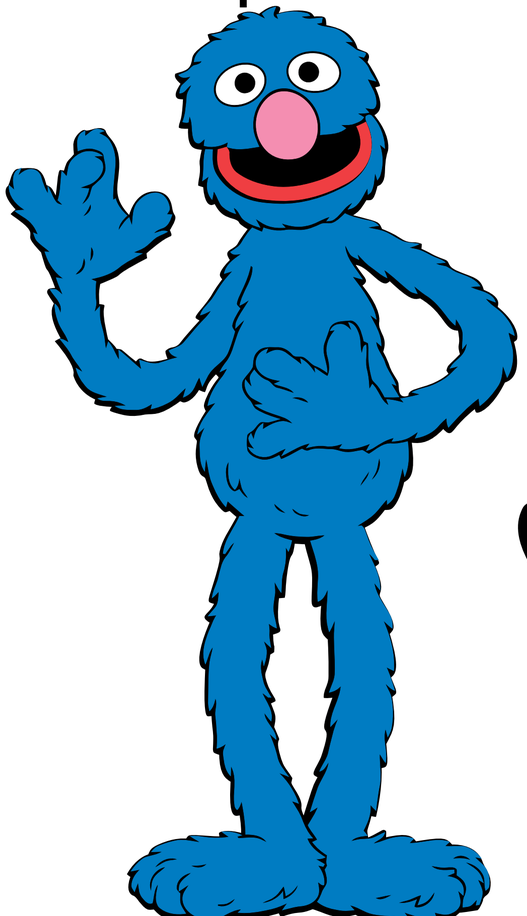


Three sizes

Grover-Base,
117M parameters

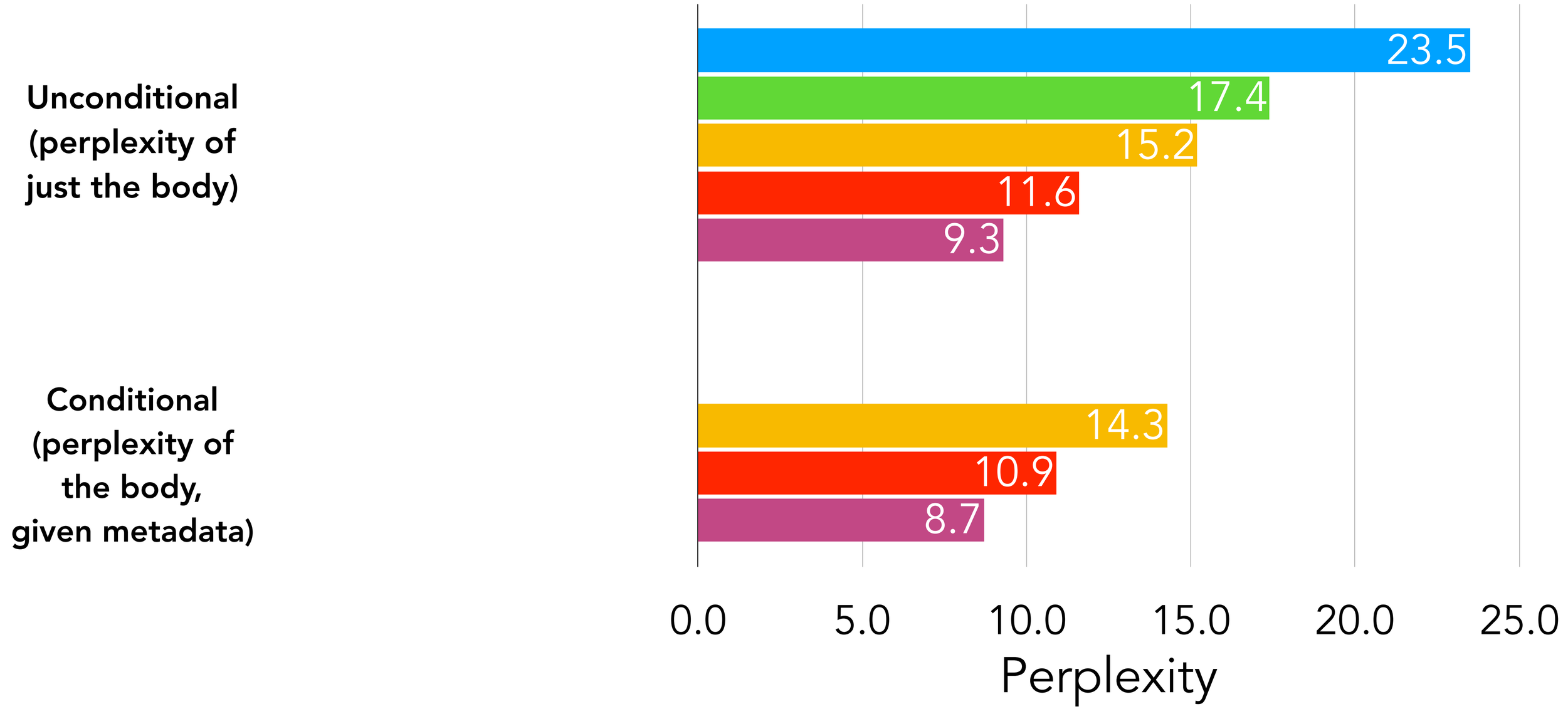


Grover-Large,
345M parameters



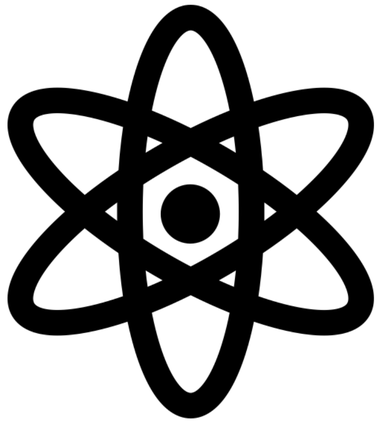
Grover-Mega,
1.5B parameters

How important is domain+metadata?



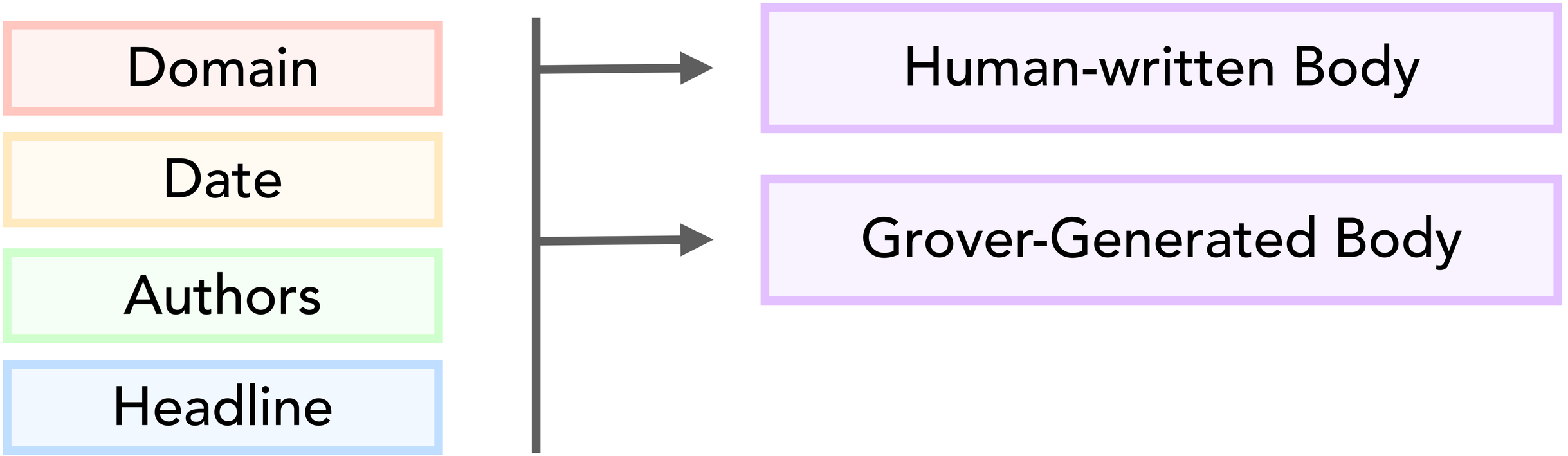
Sampling from Grover?

We use Nucleus Sampling (to be discussed later):

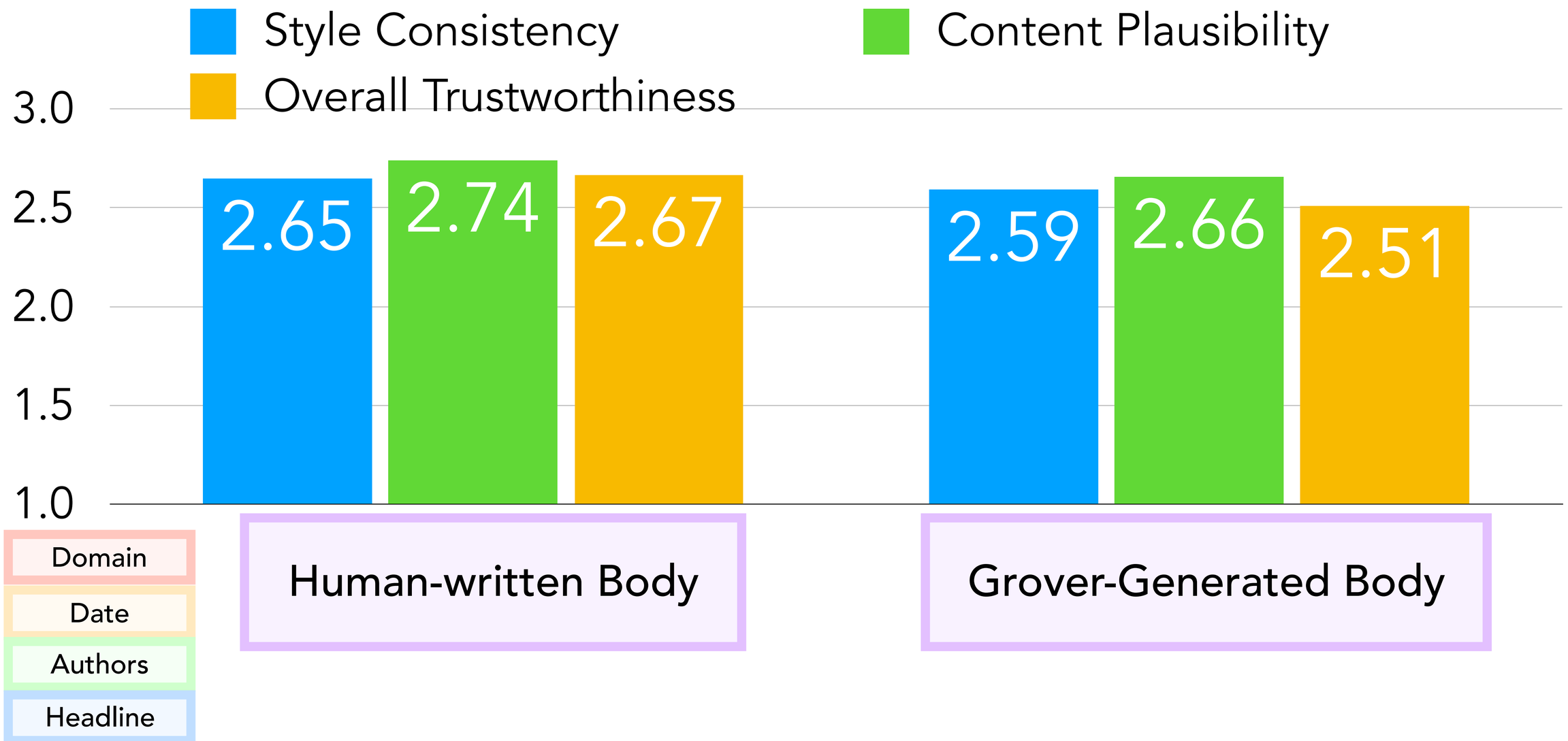


— sampling from the top ***p*** probability mass

How good are Grover's generations?



How good are Grover's generations?



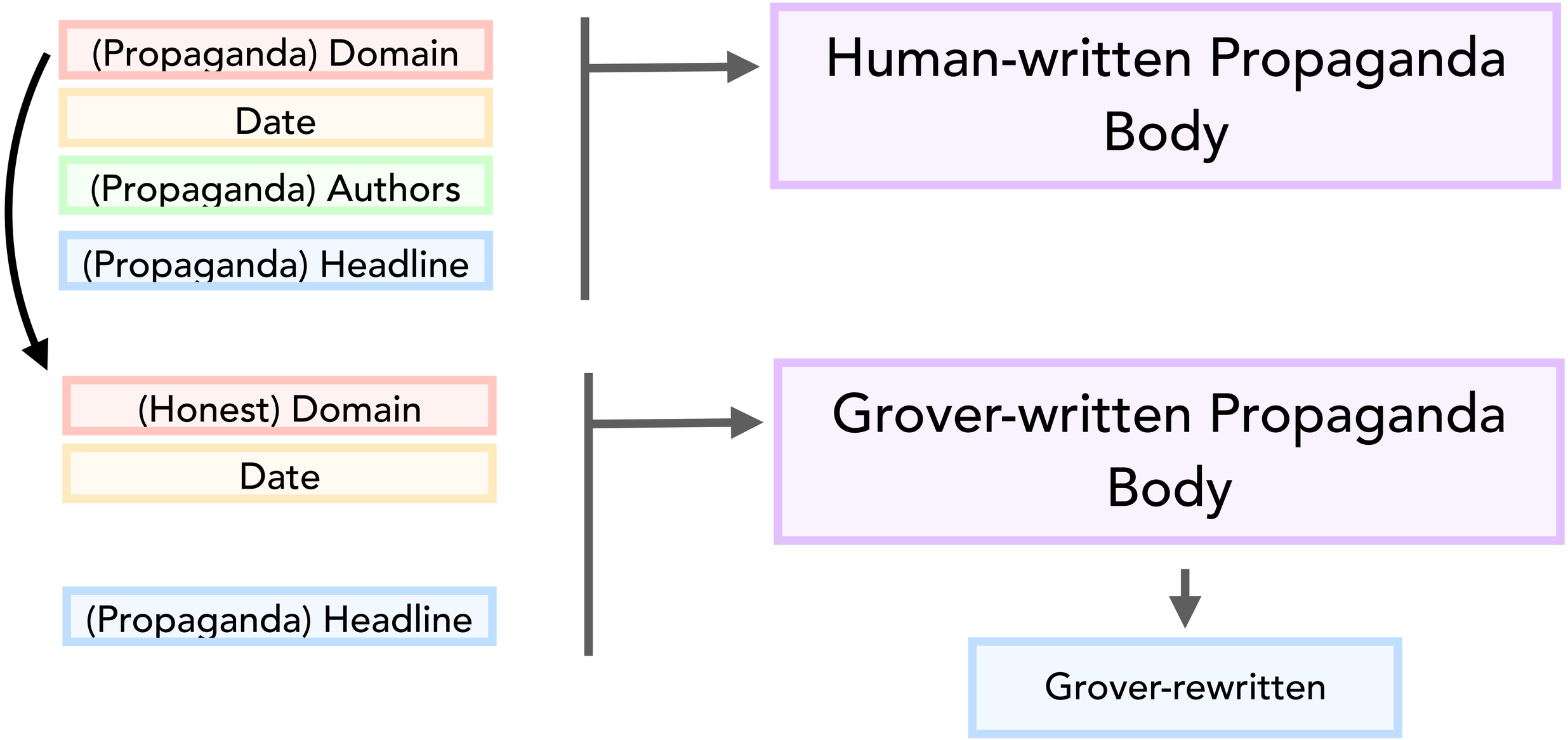
Domain

Date

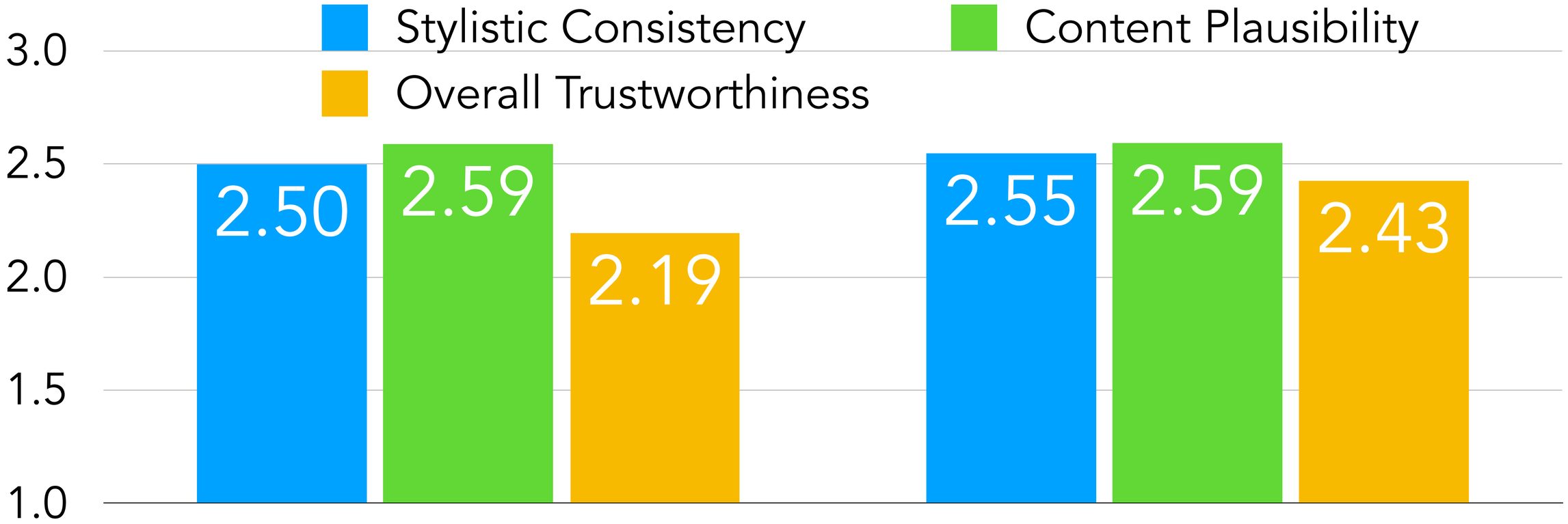
Authors

Headline

How good is Grover-written propaganda?



How good are Grover's generations?



Human Propaganda

Date

(Propaganda) Headline

Human-written

Grover Propaganda

Date

Grover-rewritten Headline

Grover-written


What have we done???

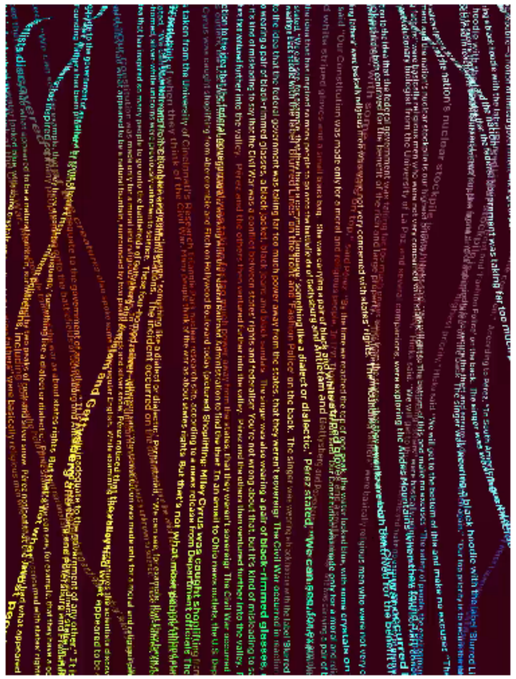
Let’s try to detect (neural) fake news!

Setup: Semi-supervised classification

- 5k examples from a Grover adversary, with a given size and generation hyperparameter ***p***
- 5k news articles published in April 2019
- Unlimited *real* news articles published earlier — used for pretraining / domain adaptation

Pretrained neural discriminators

BERT
(+ domain adaptation)



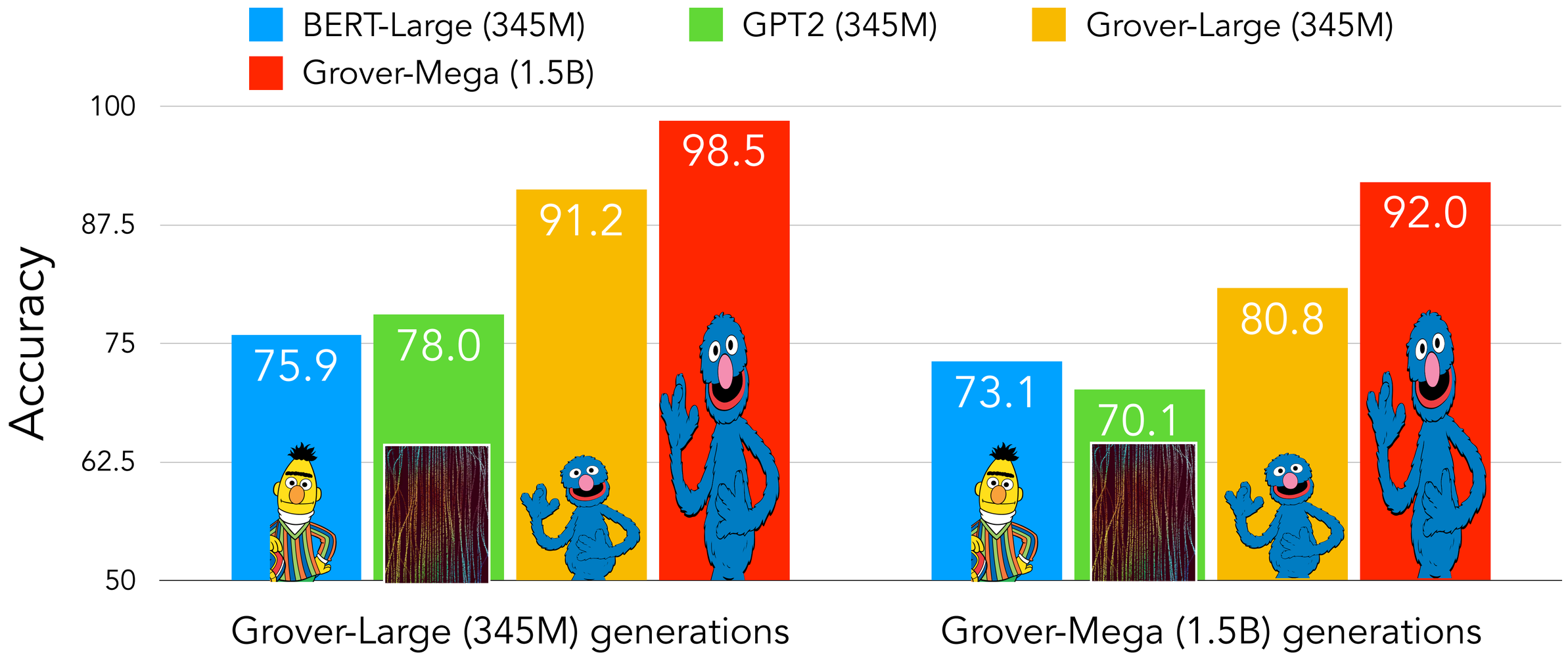
GPT2

Grover
(with a different checkpoint)

(Radford et al., 2018, Devlin et al., 2018)

Discrimination results

For each generator-discriminator pair, we use the most adversarial generation hyperparameter ***p***.



Why is Grover the best at detecting neural fake news written by other Grover models?

(this is surprising! Conventional wisdom is that the discriminator should have a different structure)

The Curious Case of Neural Text *De*generation

April 22 2019, arXiv

Ari Holtzman

Max Forbes

Jan Buys

Yejin Choi

Previously On...

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.

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.

Isn’t Beam Search Better?

BeamSearch, b=10:

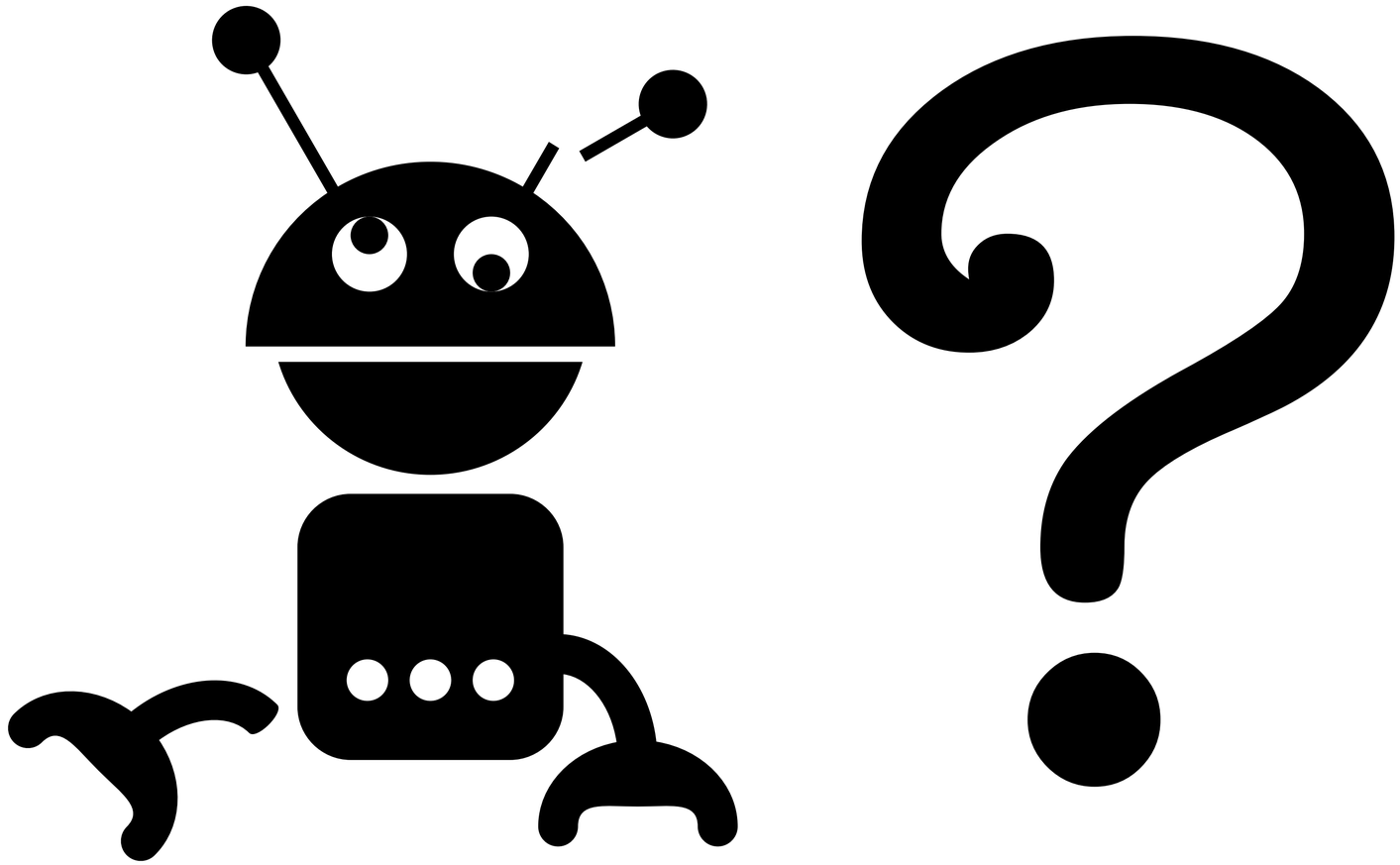
"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition

Isn’t Beam Search Better?

BeamSearch, b=10:

"The unicorns were able to communicate with each other, they said unicorns. a statement that the unicorns. Professor of the Department of Los Angeles, the most important place the world to be recognition of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the world to be a of the…


That looks like
NEURAL TEXT
DEGENERATION



The Curious Case of Neural Text Degeneration

Hypothesis #1:

Language Models are ***so*** good these days! Let’s just take their best bet at every step.

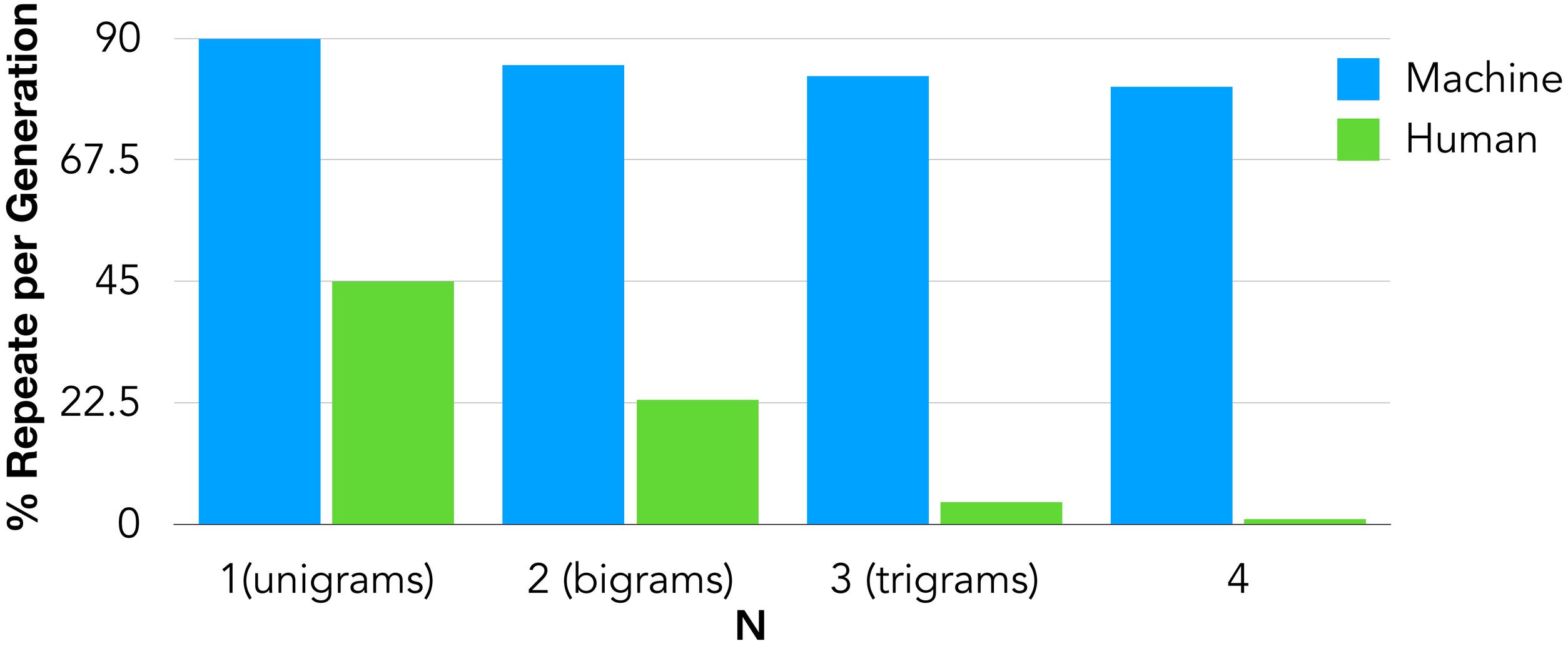
Greedy Decoding

$$\max_{w_i \in V} P(w_i | w_1 \cdots w_{i-1})$$

Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: The headline read: ”The New York Times.” The headline read: ”The New York Times.” The headline said: ”The New York Times.”

But that’s just one example!



Hypothesis #2:

Language Models are really good.

But we need to use a smarter algorithm to maximize the probability of the entire generation.

Beam Search!

$$\max_{w_i \cdots w_j \in \{V\}^{j-i}} P(w_i \cdots w_j | w_1 \cdots w_{i-1})$$

Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: The headline read: "New York City, New York, New York, New York, New York, New York, New York, New York, New York, New York, New York,

What is going on??!!!

"Just a bug"

—

Bug in the beam search code?

"Search Error"

—

Use a larger beam? (to find a sequence with higher prob?)

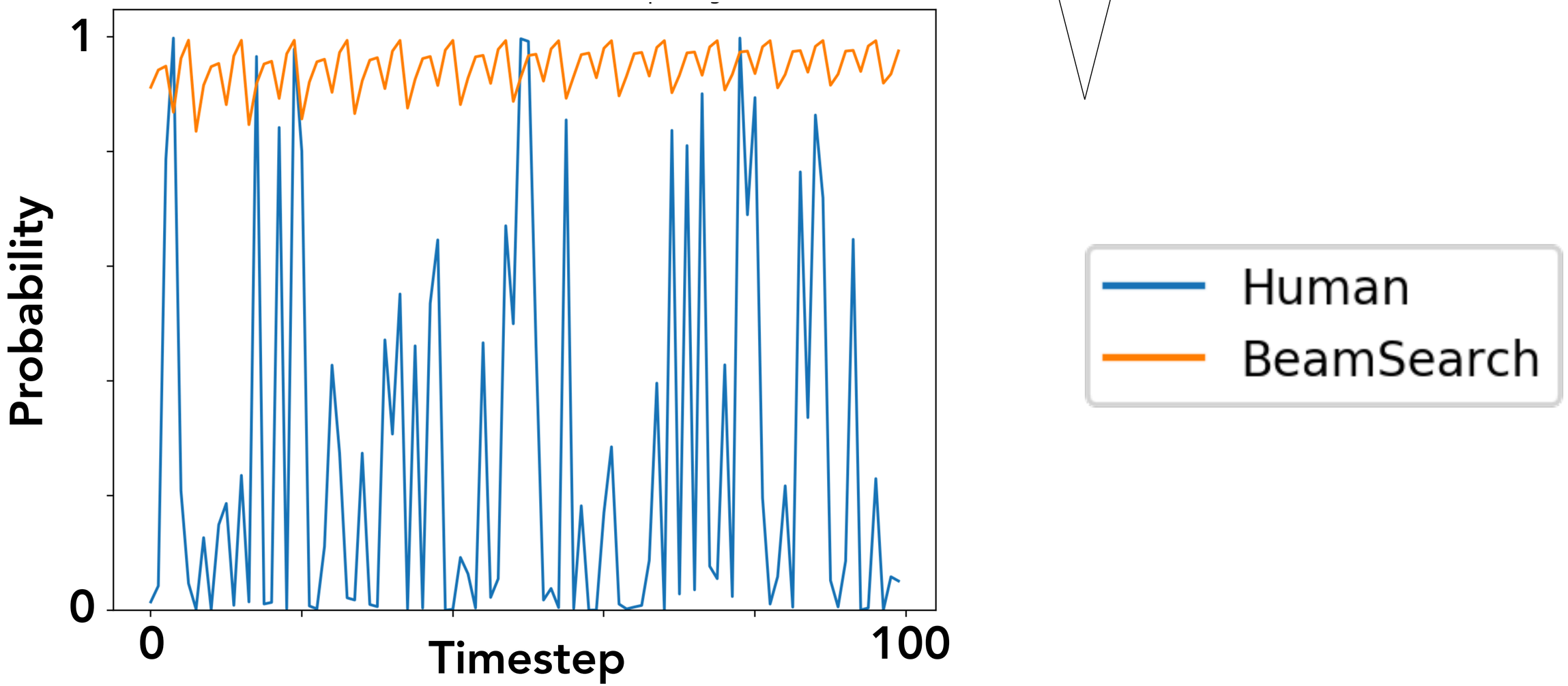
"Model Deficiency"

—

NNs not big enough? (Need GPT-3?)

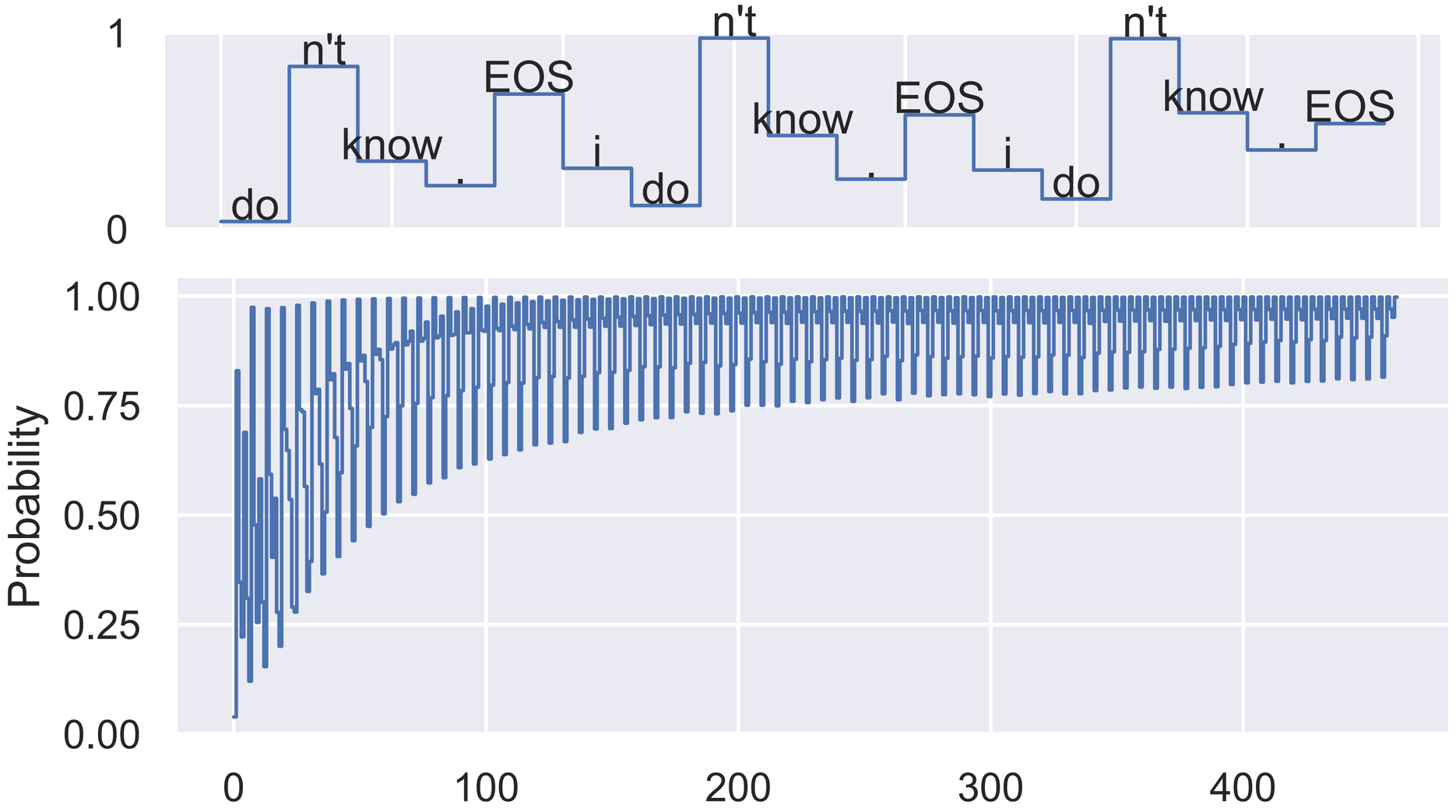
or???

The natural distribution of human text has lots of spikes.
In contrast, the distribution of machine text (based on max likelihood decoding) is artificially high and flat!



The Curious Case of

“I don’t know. I don’t know. I don’t know. …”

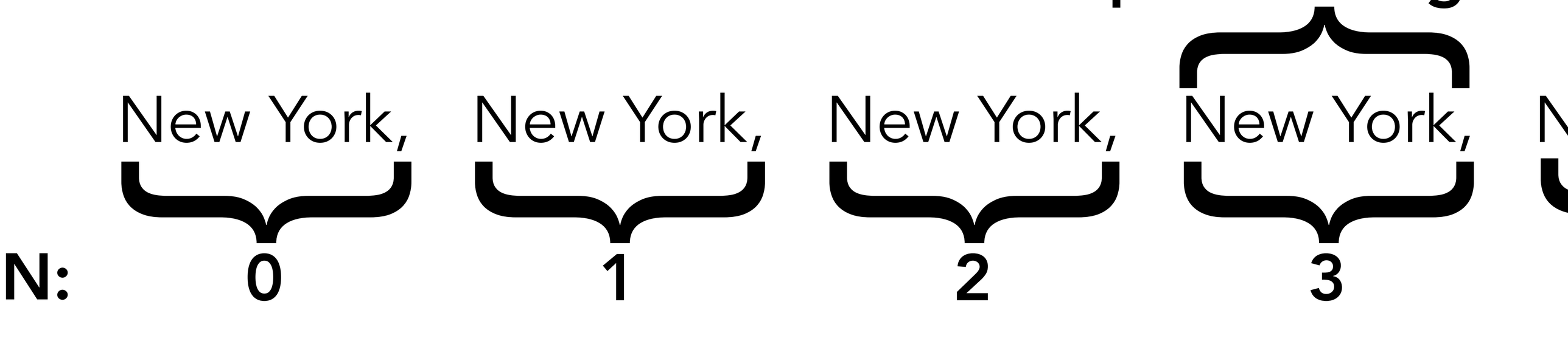


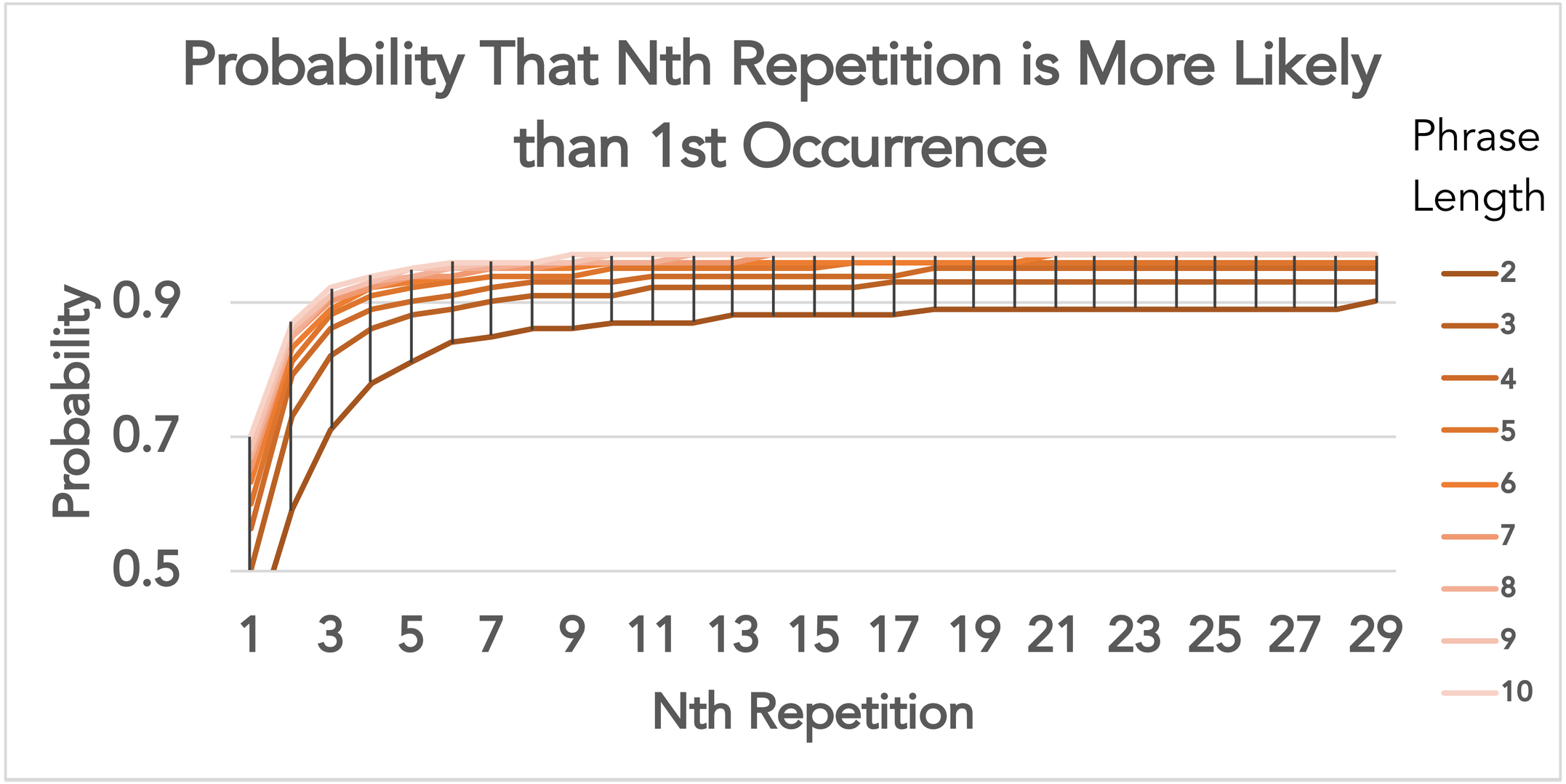
Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: The headline read: "New York City, New York, New York, New York, New York, New York, New York, New York, New York, New York, New York,

Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting. The headline read "New York City,

M=3 token phrase length




Probability That Nth Repetition is More Likely than 1st Occurrence
Phrase Length
Probability
0.9
0.7
0.5
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29
Nth Repetition
2
3
4
5
6
7
8
9
10

Why Doesn’t Maximization Work?

- Humans aren’t attempting to maximize probability, they’re trying to achieve goals. (Goodman, 2016)
- Successful language models all rely heavily on attention, which easily learns to amplify a bias towards repetition.
- Maximization is problematic in high-entropy timesteps, regardless of the quality of the language model.

Hmm…

Hypothesis #3:

Actually, we want something directly from the distribution of language.

Pure Sampling

$$w_i \sim P(w_i|w_1 \cdots w_{i-1})$$

Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: He had opened the crossword puzzle and was pointing the newspaper from it. And the title: 12:50pm how happy has white rabbit been? why is They declining white rabbit?

Question:

Why ***de***generation into gibberish?

Answer:

The ***(long) tail*** of the distribution is where the quality of LMs become worse.

Hypothesis #4:

Let’s sample from
only the head distribution
(and cut the tail distribution)!

Top-*k* Sampling

$$w_i \sim \text{best\_}k\text{\_options}\left(P(w_i|w_1 \cdots w_{i-1})\right)$$

Fan et al. 2018, Holtzman et al. 2018
Radford et al. 2019

Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: He had seen the news, but had not read the New York times or the times. The local post would have been much quicker, perhaps even better.

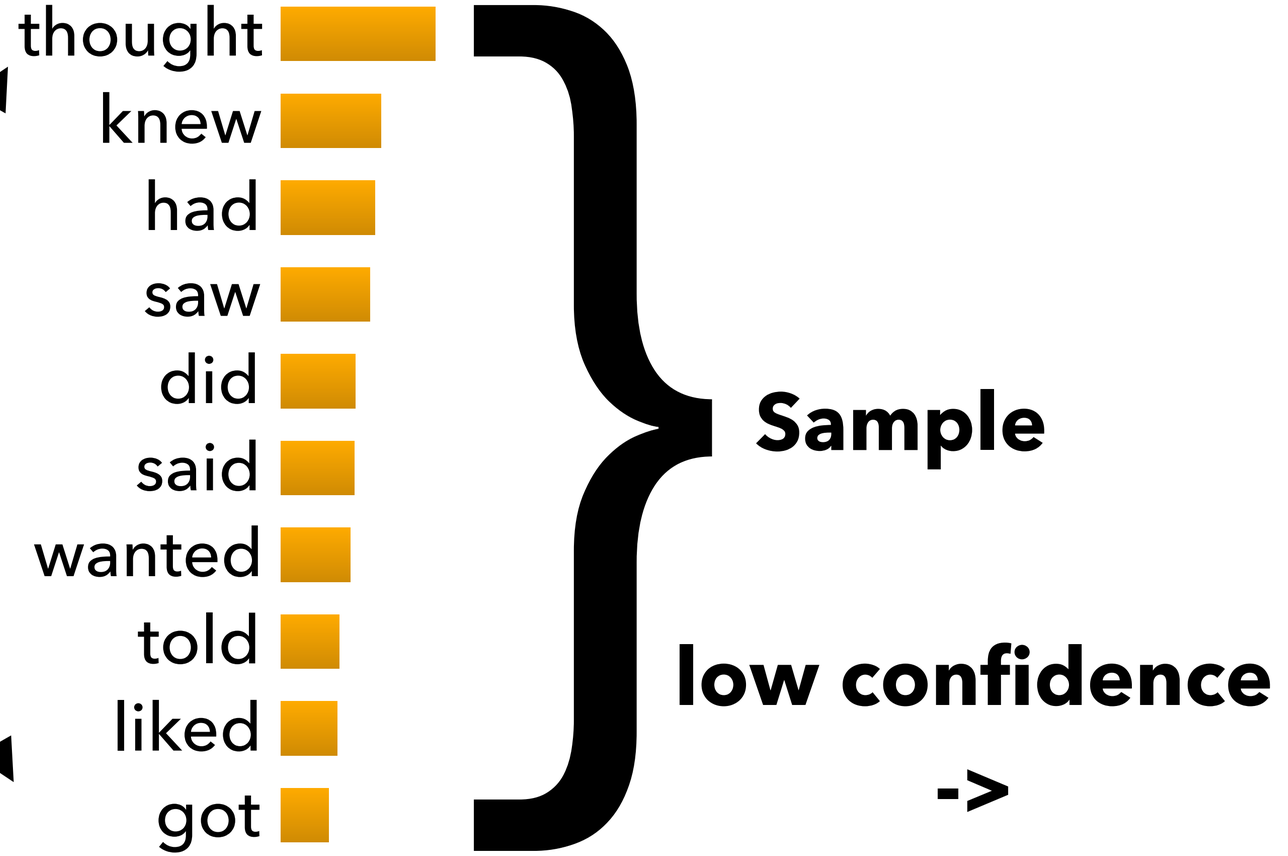
Better, but still two problems

1. top-k can be still too bland

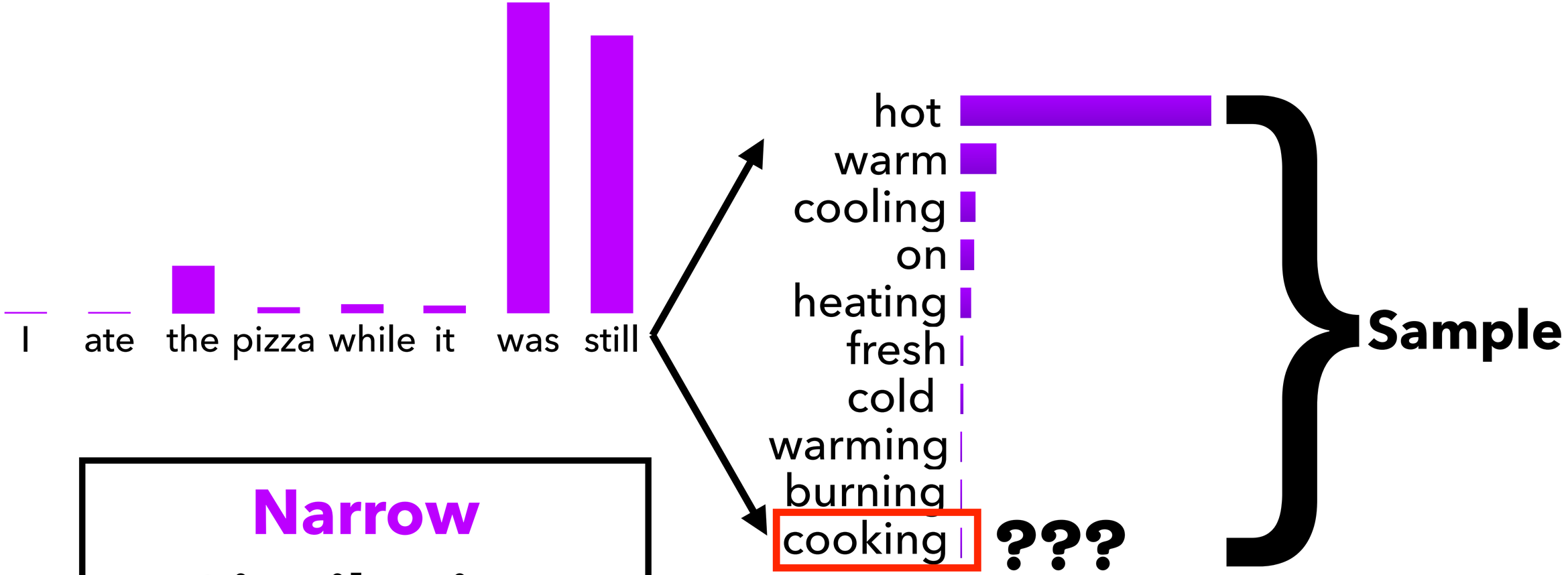
Top-*k* Sampling

Broad

Distribution



2. Top-*k* can also be too random



Narrow Distribution

high confidence

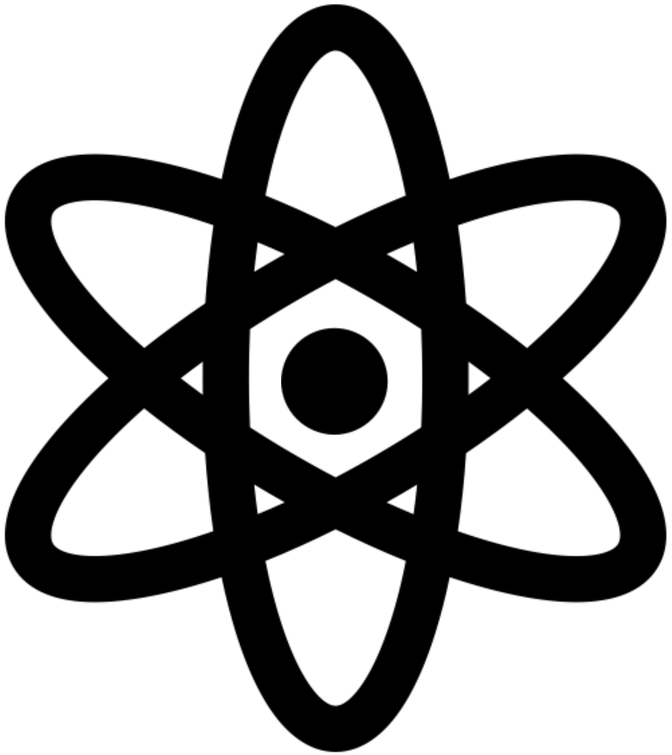
->

some extremely low probability choices

Hypothesis #5:

Most of the probability mass is in the right place, but we want to ***avoid "the dregs"*** at the bottom of the coffee.

The **Nucleus** of the Distribution



The **Nucleus** of the Distribution

- The small subset of vocabulary where the most probability mass is concentrated.
- Top-*k* fails to account for the dynamically changing per-word distributions: some very skewed, some very flat.
- Let's sample from top-p nucleus with dynamically shrinking and expanding top-k!

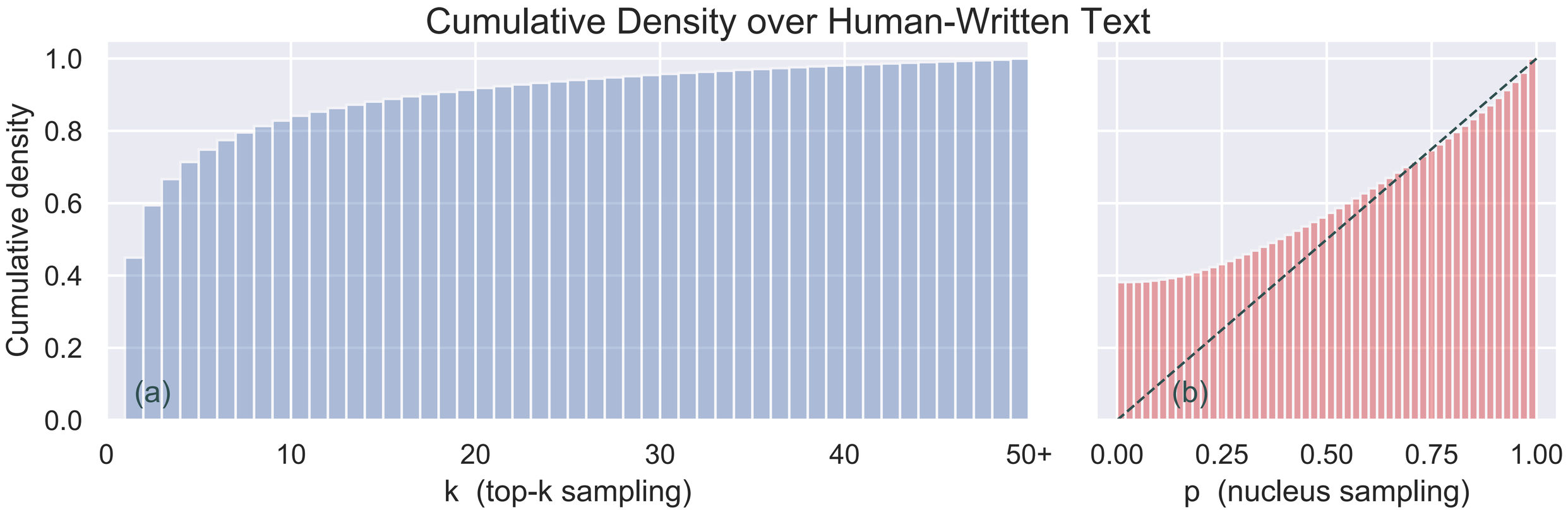
Nucleus (Top-*p*) Sampling

$$w_i \sim V^{(p)}$$

where

$$V^{(p)} = \sup_{V' \subset V} \sum_{x \in V'} P(x|w_1 \cdots w_{i-1}) \geq p$$

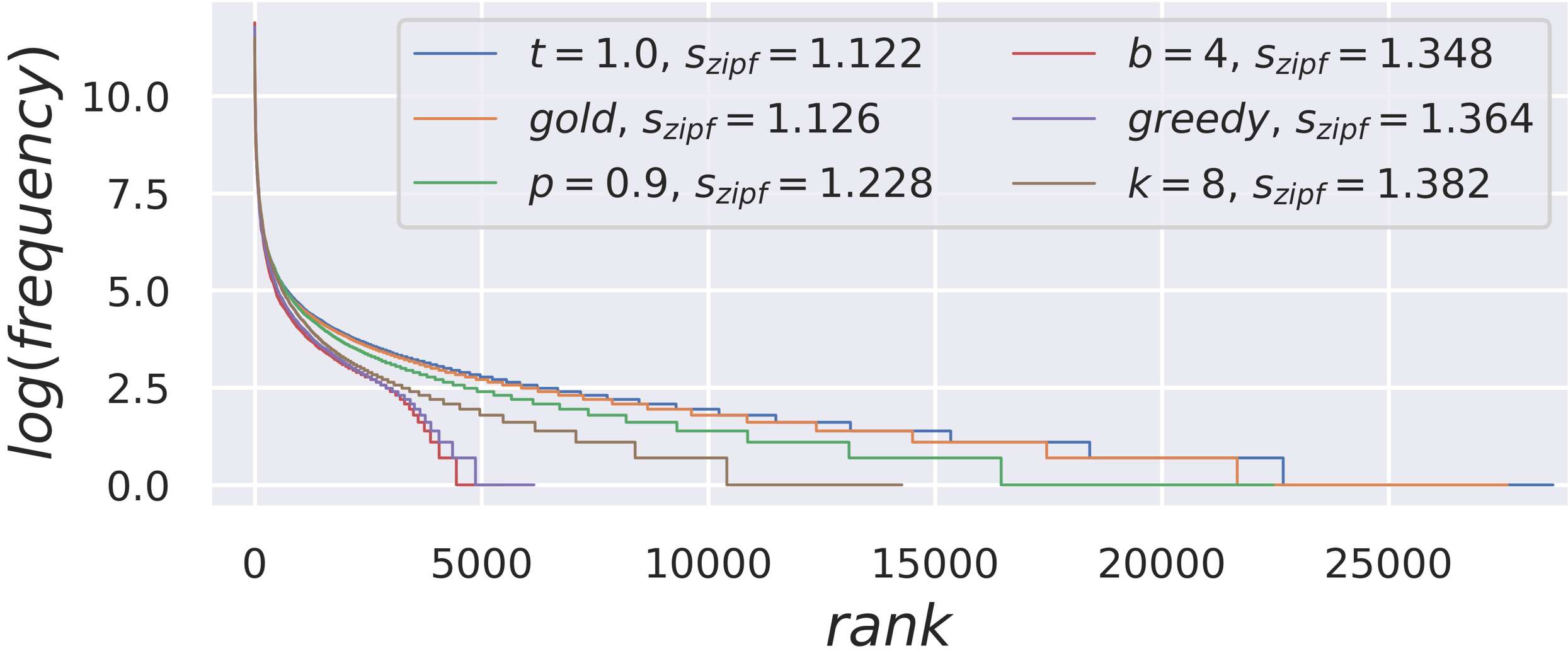
Top p more effective than Top k



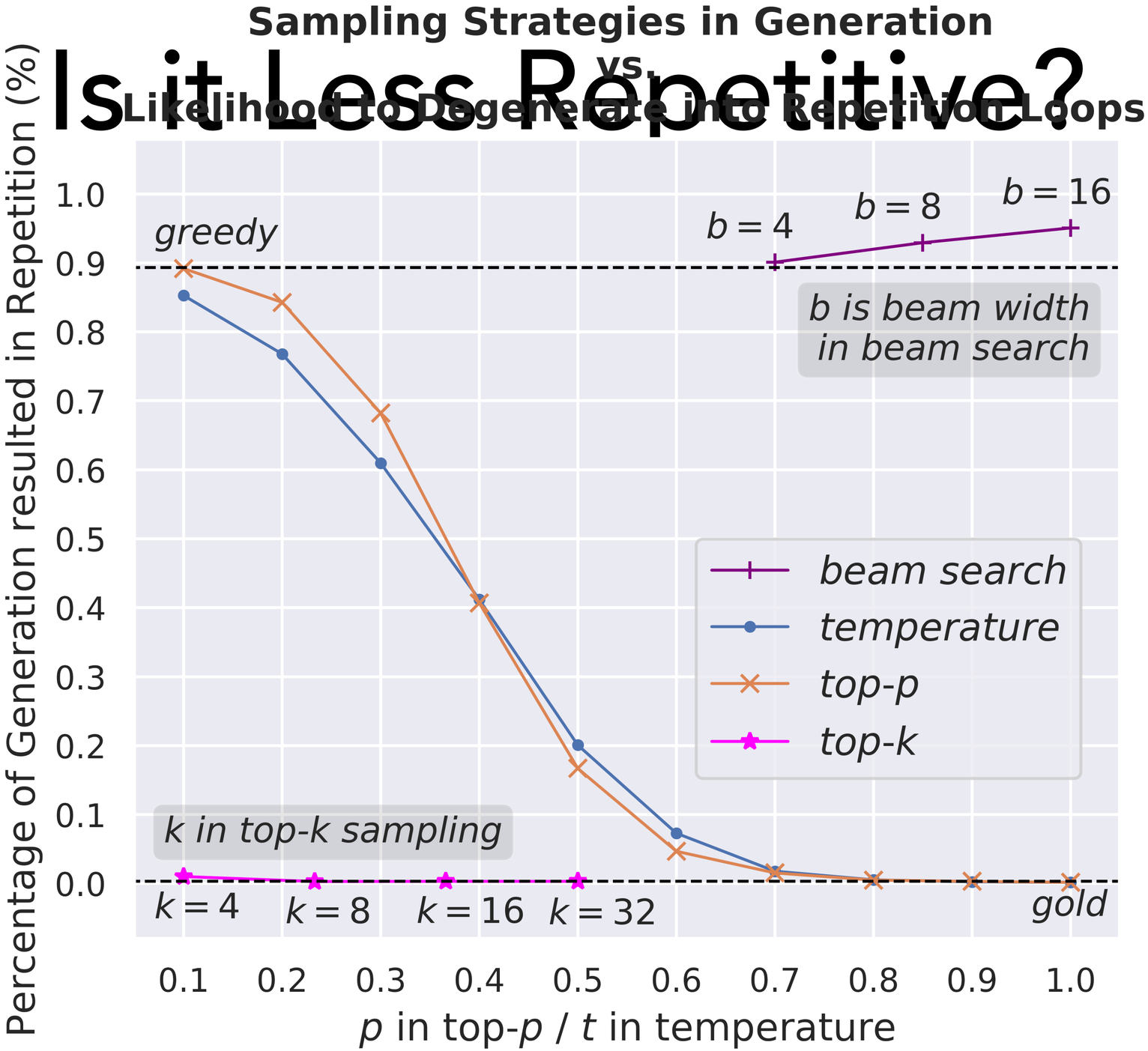
Initial: Dwight arose from his bed. He walked down stairs, He made his breakfast, and he sat at the finely crafted wooden dinner table. At his right, a cup of coffee. At his left, the news paper. The crossword puzzle was particularly interesting.

Continuation: It was on the ground floor of the Imperial Hotel. He could hear the TV from the lobby of the palace. There were headlines that would make a cop blush.

Is it Less Bland?



Is it Less Repetitive?

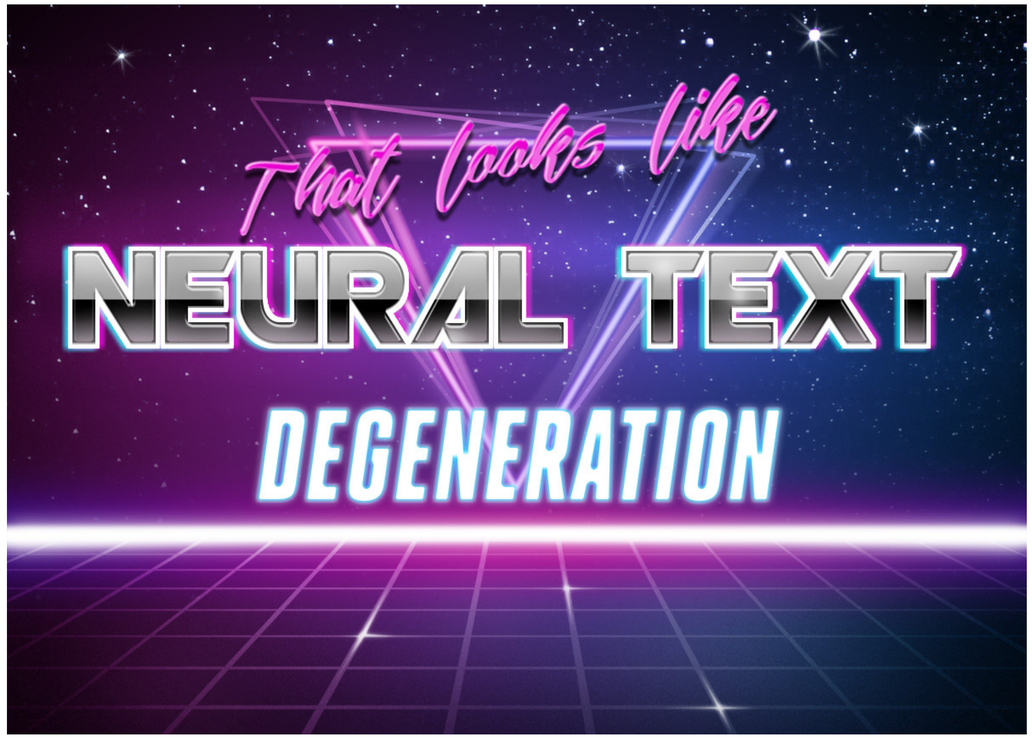


Takeaways

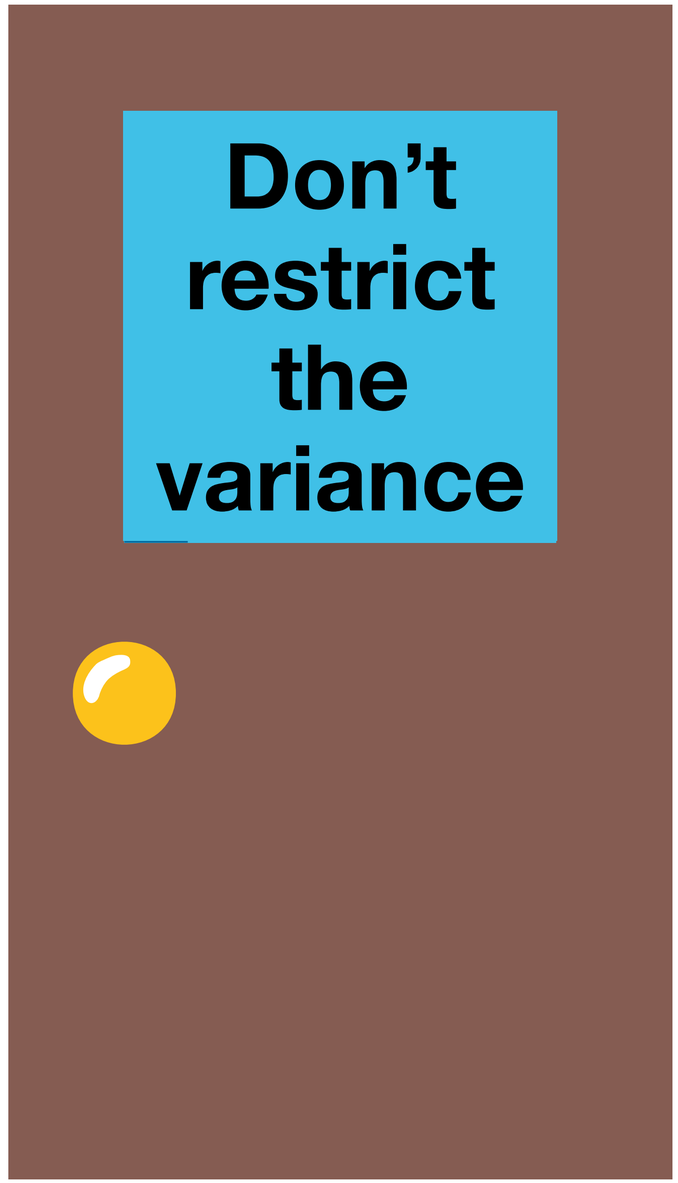
- Likelihood maximization methods (like Beam Search) leads to unnatural distribution that is too high and flat.
- Top-*k* Sampling has issues not accommodating dynamically changing shapes of the per-word distributions.
- Nucleus Sampling allows the pool of candidates to ***expand and contract dynamically.***

Back to the earlier question…

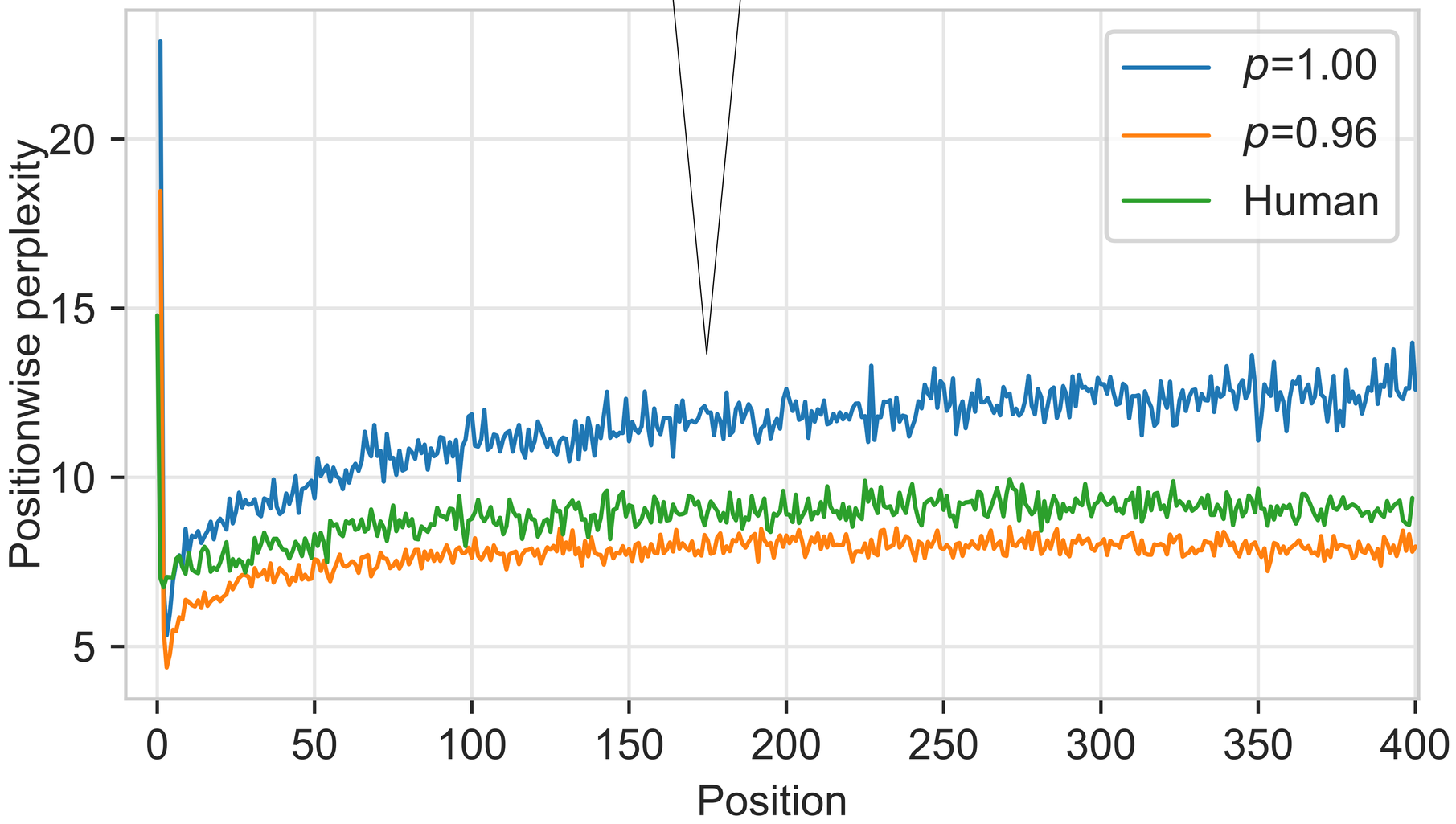
Why is Grover good at detecting neural fake news written by other Grover models?



Have to pick one of the two bad choices



A Clear Case of Neural Text ***De***generation!

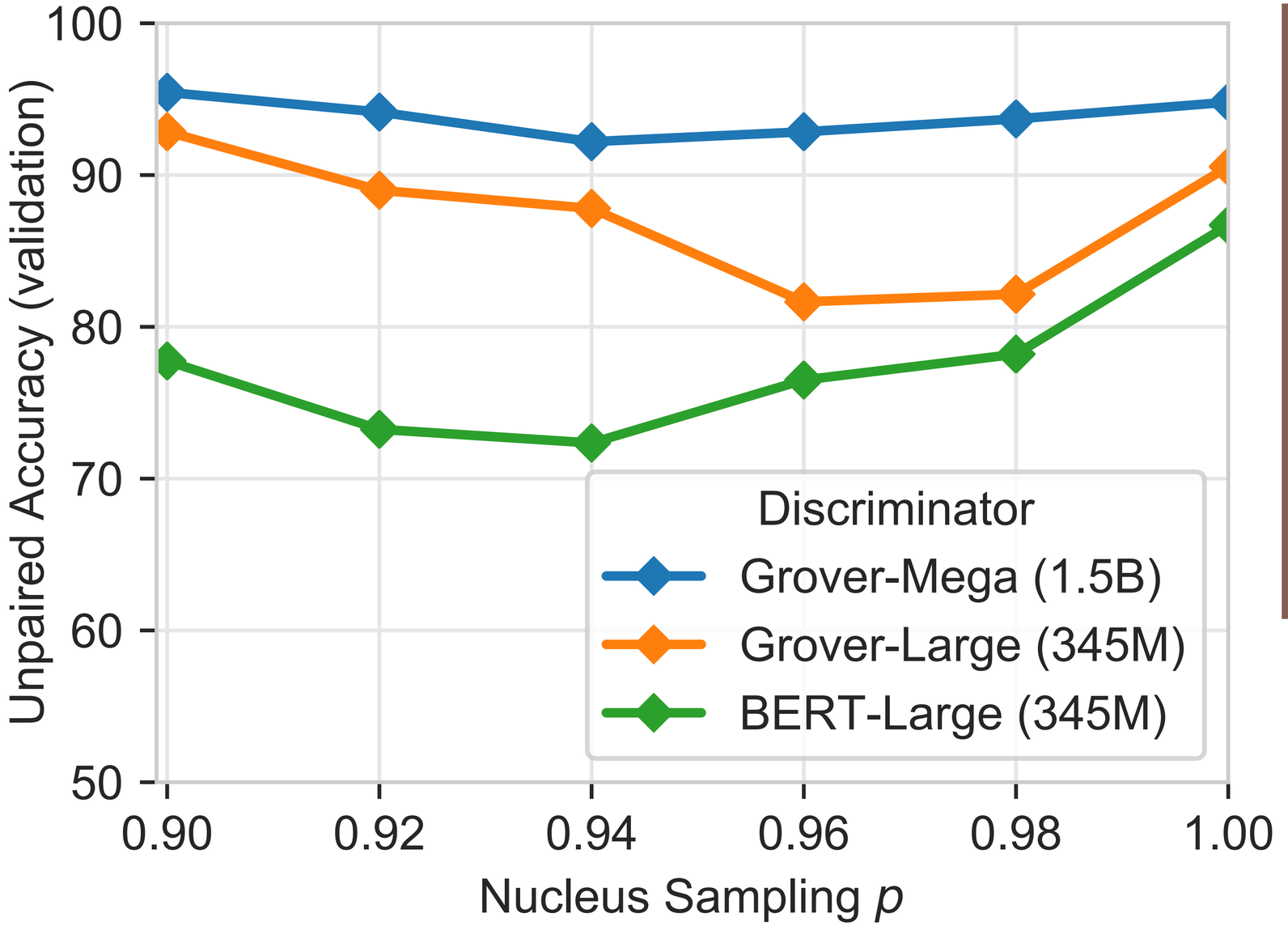


Don’t restrict the variance

Restrict the variance

- The probability that all tokens are in the top-p head is p^N, which goes to 0 as the number of tokens N gets large.
- Which leaves a distributional signature of the neural authorship!
- It is the generator itself that knows its own *"habits, quirks and traits"* the best!

Restrict the variance



Don't restrict the variance

Demo time!

rowanzellers.com/grover

https://grover.allenai.org/

Conclusion

- Modern security research builds on thread models
- The best defense against the generator is the generator itself
- Our position: Grover is not a panacea, and no AI algorithm ever will be. We need collaborative research across AI and Security research. Our study is one step toward understanding and defending malicious uses of neural language models.
- Release plan: Will share base and medium right away. For Mega, researchers can apply for it.

What would an adversary do?

Disinformation: Fake News Intended to Deceive

Ad Revenue!
(generate only viral content)

Persuade people!
(generate content that fits a worldview)

(Wardle, 2017; Bradshaw and Howard, 2017; Melford and Fagan, 2019)

Q. still feel concerned about releasing mega-sized models

- Modern security research builds on thread models
- The best defense against the generator is the generator itself
- The can of worms has already been opened (before our study)
- Neural fake news often do not serve the purpose of particular goals of fake news generation: ad revenue with viral content / propaganda, as more fine-grained control is still an open research question.
- In any case, Grover is not a panacea, and no AI algorithm ever will be. We need collaborative research across AI and Security research.

Q. What about rejection sampling?

- We tried something along that line using "**adversarial filtering**" — used for SWAG (Zellers et al. 2018), and HellaSWAG (Zellers et al 2019) — an algorithmic way to filter out easy examples and retain only hard samples.
- We find that adversarial filtering just doesn't work in this setting
- 92% accuracy does not mean Grover can actually "generate" a passable article with 8% (or 4%?) of the chance for any given context!