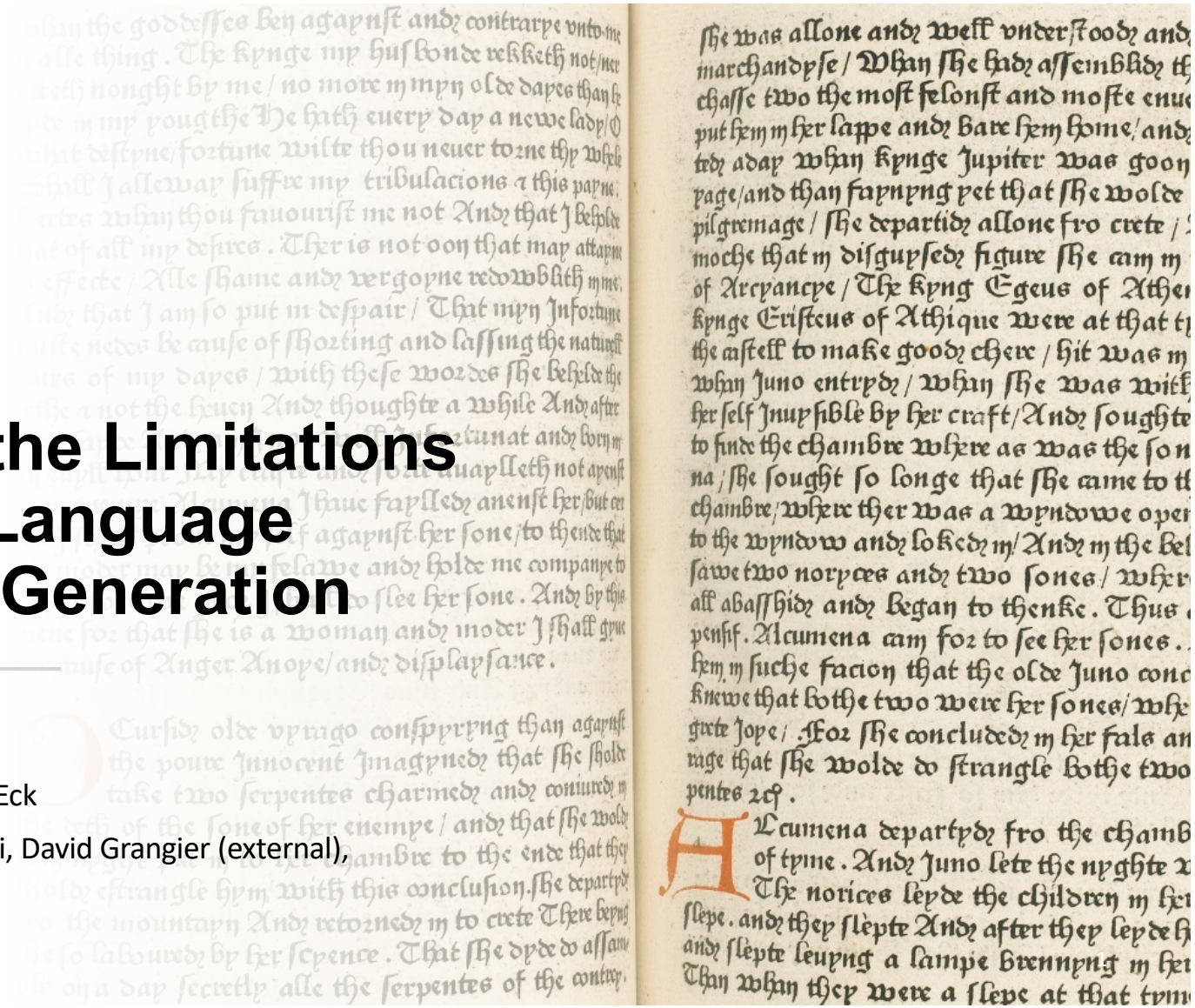
Understanding the Limitations of Using Large Language Models for Text Generation

Daphne Ippolito

Advisors: Chris Callison-Burch, Douglas Eck

Thesis committee: Marianna Apidianaki, David Grangier (external), Dan Roth, Lyle Ungar (chair)



Thesis: Understanding the Limitations of Using Large Language Models for Text Generation

Text generation encompasses tasks that involve generating word sequences in a natural language, like English.

Most often we care about *conditioned* text generation: given some input and a goal, generate a word sequence that achieves the goal.

NLG = natural language generation

Text Summarization

THE HILL

Canada tapping reserve maple syrup supply amid shortage

BY MAUREEN BRESLIN - 11/29/21 09:23 AM EST

852 SHARES SHARE TW

Just one week after President Biden tapped into oil reserves to help with a global shortage, a Canadian group has announced that it will dip into its emergency supply of maple syrup to try to keep up with global demand, according to multiple reports.

The Canadian leading trade group, Quebec Maple Syrup Producers, is releasing nearly 50 million pounds of its reserve, which is about half of the stockpile, according to CNN Business.

The trade group is backed by the Canadian government and is at times dubbed the "OPEC of Maple Syrup." As of 2020, Quebec produced 73 percent of the world's supply of maple syrup, CNN reports. By far, the largest buyer of the goods is the United States.

The reserve was created to prepare for situations, such as these, when there are poor harvest seasons or a spike in demand for maple syrup, reports CNN.

NPR reports that this year saw uncharacteristic warmth in a short spring, which gave maple syrup harvesters little to work with.

"That's why the reserve is made, to never miss maple syrup. And we won't miss maple syrup!" said Helene Normandin, the Quebec Maple Syrup Producers' communications director when speaking with NPR.

Generated Summary:

The Canadian leading trade group, Quebec Maple Syrup Producers, is releasing nearly 50 million pounds of its reserve. The reserve was created to prepare for situations, such as poor harvest seasons or a spike in demand for maple syrup. This year saw uncharacteristic warmth in a short spring, which gave maple syrup harvesters little to work with.

Jingqing Zhang, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In PMLR, 2020.

Text Summarization

THE HILL

Canada tapping reserve maple syrup supply amid shortage

BY MAUREEN BRESLIN - 11/29/21 09:23 AM EST

852 SHARES

Just one week after President Biden tapped into oil reserves to help with a global shortage, a Canadian group has announced that it will dip into its emergency supply of maple syrup to try to keep up with global demand, according to multiple reports.

The Canadian leading trade group, Quebec Maple Syrup Producers, is releasing nearly 50 million pounds of its reserve, which is about half of the stockpile, according to CNN Business.

The trade group is backed by the Canadian government and is at times dubbed the "OPEC of Maple Syrup." As of 2020, Quebec produced 73 percent of the world's supply of maple syrup, CNN reports. By far, the largest buyer of the goods is the United States.

The reserve was created to prepare for situations, such as these, when there are poor harvest seasons or a spike in demand for maple syrup, reports CNN.

NPR reports that this year saw uncharacteristic warmth in a short spring, which gave maple syrup harvesters little to work with.

"That's why the reserve is made, to never miss maple syrup. And we won't miss maple syrup!" said Helene Normandin, the Quebec Maple Syrup Producers' communications director when speaking with NPR.

Generated Summary:

The Canadian leading trade group, Quebec Maple Syrup Producers, is releasing nearly 50 mi… reser… created to pr… s, such as po… or a spike… e syrup. This year saw uncharacteristic warmth in a short spring, which gave maple syrup harvesters little to work with.

✅ **Fluent**

✅ **Accurate**

Jingqing Zhang, et al. "Pegasus: Pre-training with extracted gap-sentences for abstractive summarization." In PMLR, 2020.

Style Transfer

Input sentence:

I had a protruding snout, rows of teeth visible in my smile.

Rewrite the input sentence to...

be more comical:

My snout was covered in warts and my teeth were crooked, but I still thought I was the most adorable dog at the doggie party!

be more descriptive:

I looked like an alligator when I smiled.. My gums were red and inflamed and my rows of teeth stuck out, making me look even more like a beast.

include a metaphor:

My smile could be compared to a shark's open mouth, rows of teeth protruding from my gums as I grinned at the people in the market.

Emily Reif*, Daphne Ippolito*, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. "A recipe for arbitrary text style transfer with large language models." In ACL 2022. See https://bit.ly/3fLDuci for more examples.

Style Transfer

Input sentence:

I had a protruding snout, rows of teeth visible in my smile.

Rewrite the input sentence to...

In the past: limited to specific style transfers we had training data for

Today: perform any arbitrary style transfer

doggie party!

beast.

people in the market.

Emily Reif*, Daphne Ippolito*, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. "A recipe for arbitrary text style transfer with large language models." In ACL 2022. See https://bit.ly/3fLDuci for more generations.

Thesis: Understanding the Limitations of Using Large Language Models for Text Generation

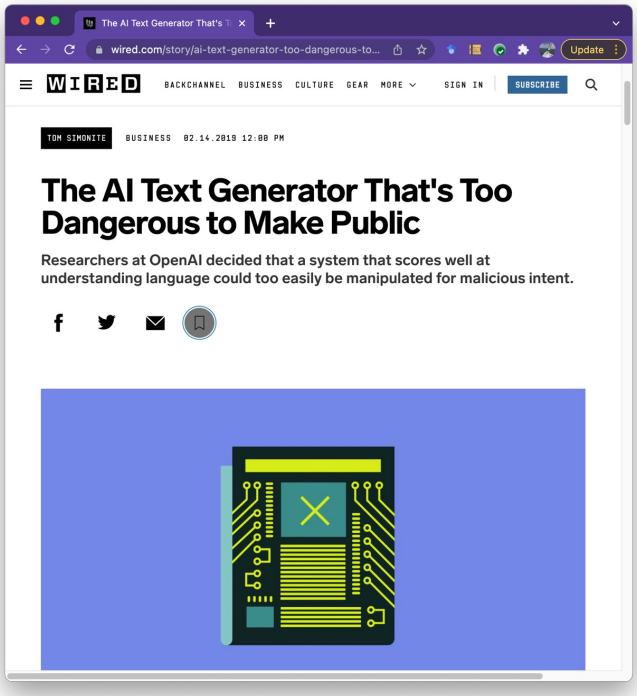
Modern language models are neural networks which have been "trained" on millions of examples of human writing.

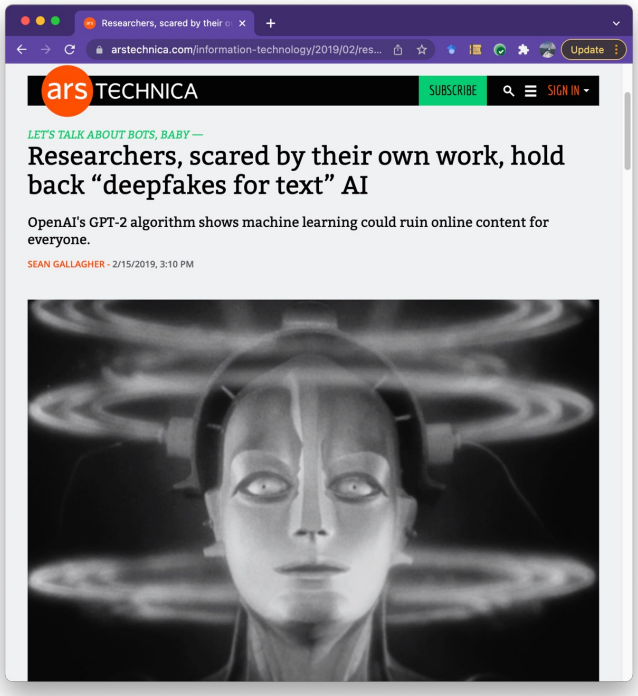
This allows them to predict text sequences that sound human-like.

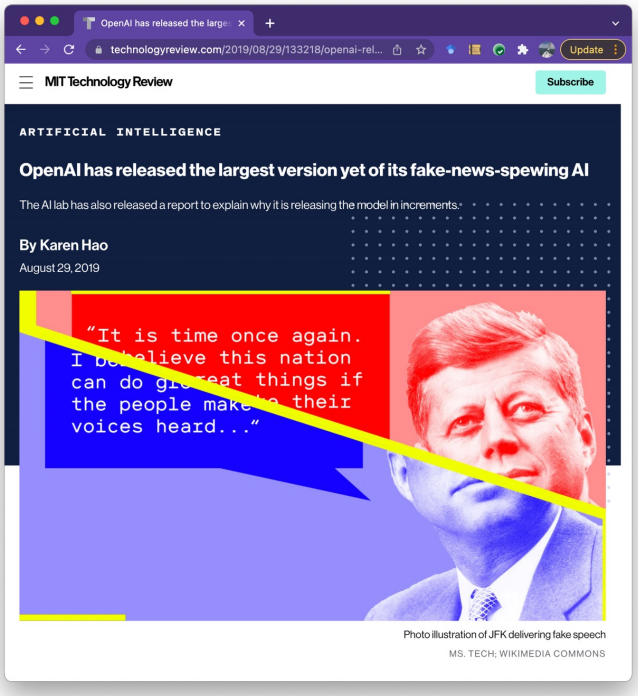
Neural networks contain "learned" parameters which get updated during training.

Models with more parameters are generally better than ones with fewer.

Thesis: Understanding the Limitations of Using Large Language Models for Text Generation

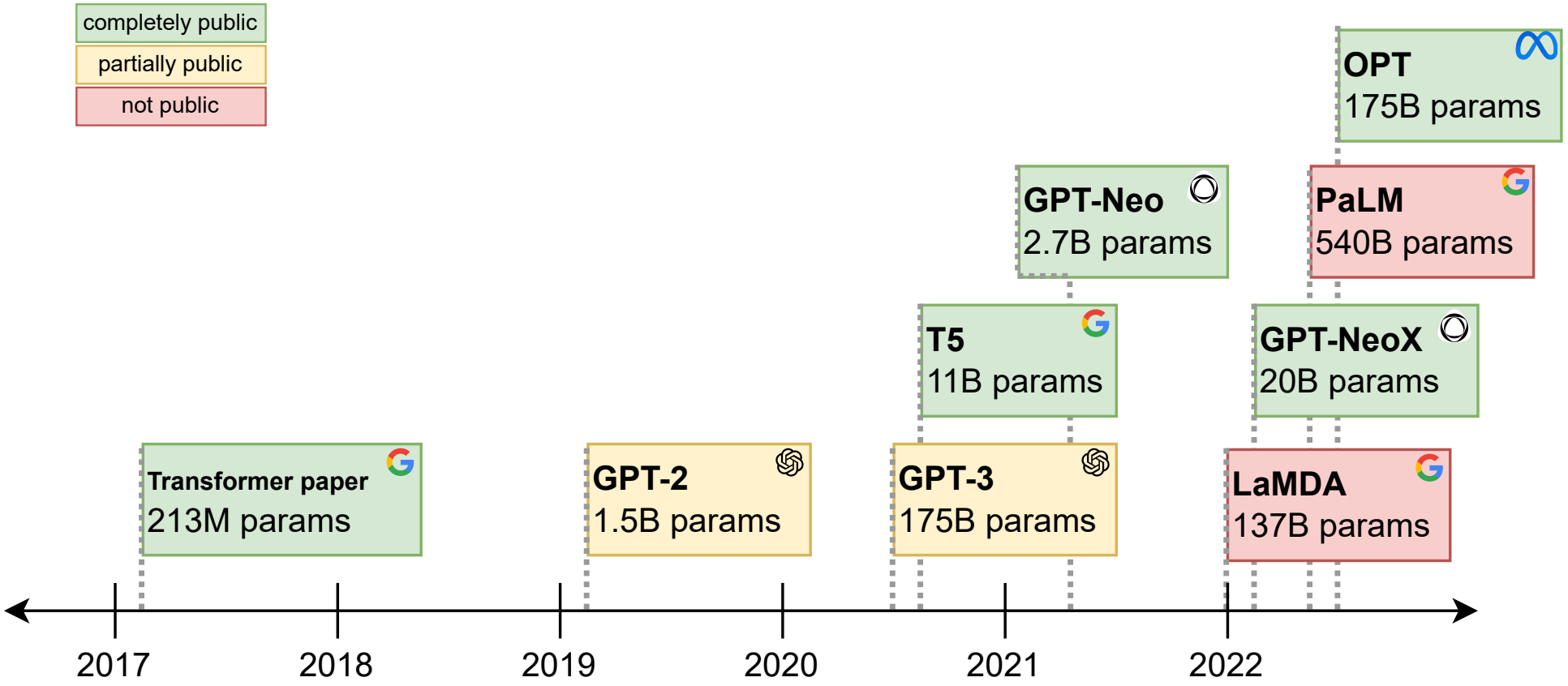






In 2019, the company OpenAI announced GPT-2 (General Purpose Transformer 2), a 1.5B parameter neural language model "too dangerous" to release.

Models Continue to Get Bigger



What are the implications?

Potential for positive impact

- Advancing research
  - NLP, robotics, and even the humanities and social sciences.
- Entertainment
  - Interactive fiction, video games, creative writing
- Human-AI collaboration for accomplishing tasks

Potential for harm

- Training data problems
  - Biased or incomplete data
  - Copyright and privacy issues
- Negative societal impact
  - Proliferation of fake news, social media posts, and reviews
  - Deep fakes that fool humans
- Expensive to train
  - Both in terms of dollars and energy usage

Thesis Statement

- We must build a deeper understanding of neural network-powered language generation systems before they are safe to deploy widely.
- This involves understanding the ways machine-generated text differs from the text a human would write given the same writing task.
  - Humans and NLG systems make different word choices.
  - NLG systems may plagiarize verbatim from their training data when asked to produce novel content.
- Differences in how humans and NLG systems write make evaluation tricky.
- NLG systems should be evaluated holistically as part of larger tools meant to assist human writers in tasks they wish to do.

Thesis Outline

1. Detecting Machine-Generated Text
2. Measuring and Mitigating the Risk of Memorization
3. Enabling Applications in Creative Writing

Talk Outline

1. Detecting Machine-Generated Text
2. Measuring and Mitigating the Risk of Memorization
3. Enabling Applications in Creative Writing

Enabling Applications in Creative Writing

Ann Yuan, Andy Coenen, Emily Reif, Daphne Ippolito. “Wordcraft: Story Writing with Large Language Models.” In IUI 2022.

Daphne Ippolito, et al. “A Recipe for Arbitrary Text Style Transfer with Large Language Models.” In *ACL* 2022.

Daphne Ippolito, et al. ”The Case for a Single Model that can Both Generation Continuations and Fill in the Blank." In submission 2022.

Daphne Ippolito, et al. “Unsupervised Hierarchical Story Infilling” In *ACL Workshop on Narrative Understanding* 2019.

Measuring and Mitigating the Risk of Memorization

Daphne Ippolito, et. al. “Deduplicating Training Data Makes Language Models Better.” In *ACL* 2022.

Chiyuan Zhang, Daphne Ippolito, et al. “Counterfactual Memorization in Neural Language Models.” In submission 2022.

Nicholas Carlini, Daphne Ippolito, et. al. “Quantifying Memorization across Neural Language Models.” In submission 2022.

Detecting Machine-Generated Text

Daphne Ippolito, et al. "Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text." In submission 2022.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, et. al., "Trading Off Diversity and Quality in Natural Language Generation." In *EACL HumEval Workshop* 2021.

Daphne Ippolito, et al. "RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text." In *EMNLP System Demonstrations* 2020.

Daphne Ippolito, et al. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

Daphne Ippolito, et. al. "Comparison of Diverse Decoding Methods from Conditional Language Models." In *ACL* 2019.

Enabling Applications in Creative Writing

Detecting Machine-Generated Text

Measuring and Mitigating Memorization

Thesis Takeaways

Why work on creativity?

- Stories are a challenging domain.
  - Narrative arcs, dialog, character tracking, world knowledge.
- However, mistakes that are a liability in fact-based tasks can be an asset in creative ones.
- Creative writers are a great target audience.

I will tell you a tale about Nero the cat. Nero was a cat who enjoyed life. He had a bed by the window, good food, and the full adoration of his owners. But then, one day his owners came home with a dog. The dog was a slobbery, giant thing of a dog named George. The big, ugly beast snarled and bit at anything that got too close to it, including Nero. Nero was certain that his life was over, and he hated this new change. But as it turns out, Nero was wrong. the dog would save his life. One day a raccoon broke into the house.The dog stood in front of him, blocking the raccoon from getting to Nero, and snarling as he snapped at the raccoon. At first, Nero wasn't quite pleased with his dog, but as time went on, he began to look upon his dog as a sort of brother to him, a protector. From that day on, he loved his dog George, despite his slobbering, ugly ways. After that, the dog no longer snarled and bit at Nero. He loved him, because after this, the dog wasn't as scary and no longer threatened him. He now looked at the dog as a sort of family member, and felt safe by him.

181 words

Commits

[redacted]

↑ ↓ to cycle through choices (2/14)

He no longer saw him as the ugly beast from before, but as his protector and his brother. He knew that if anything were to happen to him, his dog would do anything within his power to save him, and in turn, he would go to great lengths to save his faithful companion.

He loved him, because after this, the dog wasn't as scary and no longer threatened him. He now looked at the dog as a sort of family member, and felt safe by him.

He loved him for saving him, and he loved him even when he was sloppy and drooly, even when he was ugly- - he loved him for his big protective heart and his loving eyes.

He loved him because he protected him. Nero, was now safe from the world and the people in it, from those who hurt him, and from those who didn't care who he was or what he'd been through. Because of that, he didn't care [redacted] very ugly and big, because he saw through his appearance and saw a protector, a comrade,

Controls in Wordcraft

- **Story seeds**: When the editor is empty, the system proposes story ideas according to a user-specified prompt
- **Continuation**: The system proposes a continuation to the text to the left of the user's cursor.
- **Elaborate selection**: The text the user has selected is expanded upon.
- **Generate text from a custom prompt**: The user can type in a prompt, for example “tell me what the cat does next” and get suggestions.
- **Fill in the blank**: The user sees alternatives for the text they have selected.
- **Rewrite selection**: The user can ask Wordcraft to rewrite their selected text according to some property. For example, "to be Shakepearean" or "to include a metaphor.”

Controls in Wordcraft

- **Story seeds**: When the editor is empty, the system proposes story ideas according to a user-specified prompt
- **Continuation**: The system proposes a continuation to the text to the left of the user's cursor.
- **Elaborate selection**: The text the user has selected is expanded upon.
- **Generate text from a custom prompt**: The user can type in a prompt, for example “tell me what the cat does next” and get suggestions.
- **Fill in the blank**: The user sees alternatives for the text they have selected.
- **Rewrite selection**: The user can ask Wordcraft to rewrite their selected text according to some property. For example, "to be Shakepearean" or "to include a metaphor.”

All controls use LaMDA as the underlying language model.

Wordcraft User Study: Novice Writers

- **Participants**: 25 novice writers recruited from a mailing list for creative writing
- **Goal**: Understand whether users preferred the full Wordcraft controls to:
  a) A text editor that just suggests continuations
  b) A text editor alongside a conversational chatbot
- **Task**: Writers spend 10 minutes writing a story in each experimental setting, following a provided prompt.
- **Takeaways**:
  - Participants found the full Wordcraft setting to be more helpful than the two baseline settings.
  - Participants appreciated the chatbot interface even though it couldn't "read" their stories.
- **Study Limitations:**
  - Controlled experimental setting
  - All participants were Google employees

Controls in ⭐Updated⭐ Wordcraft

- **Story seeds**: When the editor is empty, the system proposes story ideas according to a user-specified prompt
- **Continuation**: The system proposes a continuation to the text to the left of the user's cursor.
- **Elaborate selection**: The text the user has selected is expanded upon.
- **Generate text from a custom prompt**: The user can type in a prompt, for example “tell me what the cat does next” and get suggestions.
- **Fill in the blank**: The user sees alternatives for the text they have selected.
- **Rewrite selection**: The user can ask Wordcraft to rewrite their selected text according to some property. For example, "to be Shakepearean" or "to include a metaphor.”

- **Chatbot interface**: The user can converse with a chatbot about their story.

Wordcraft User Study: Experienced Writers

- **Participants**: 13 published authors, including poets, novelists, script writers.
- **Goal**: Understand the potential and limitations of NLG technology for writers with established styles and workflows.
- **Tasks**: Write a ~1,500 word story for a digital literary magazine.
  Keep a journal of their experience.

How did writers successfully use Wordcraft?

- Idea Generator
  - "a place on my computer to go that looked like a blank page but did not behave as such"
- Brainstorming Partner
  - “The chat is an amazing tool for brainstorming or rubber-ducking. Its conversational quality is perfect to talk about plot, characters and wordbuilding.”
- Research assistant
  - "It’s kind of great to use the chat interface and treat LaMDA as a thesaurus, quote finder, and general research assistant.”
  - Asking Wordcraft: “how would a robot argue this point?”
- Generation of specific details
  - "I asked Wordcraft to come up with some names for some “Maggot Warriors” and then turn these names into bios and descriptions."
- Explorations of Rewriting
  - Rewriting the same passage many different ways.

How did the system fail?

- Struggled to capture distinct styles and voices
  - Only one consistent writing style: “novice fan-fiction writer”
  - Could not handle stories with multiple distinct voices
- Preferred to "tell" rather than "show”
- Suggestions were cliché and uninteresting
- Potential for accidental plagiarism
- Suggestions got worse as the story got longer

How did the chatbot fail?

- It overpromised beyond its capabilities
  - “I'll email you a draft with my suggestions.”
- It couldn’t meaningfully answer questions about the story
  - "Which parts of the story need more details?"
  - "Does this feel like a fast-paced action story?"

Enabling Applications in Creative Writing

Detecting Machine-Generated Text

Measuring and Mitigating Memorization

Thesis Takeaways

The Detection Task

Given a passage of text, decide whether it was written by a human or generated by a natural language generation system.

How good are humans at this task?

How good are automatic discriminators?

Daphne Ippolito*, Daniel Duckworth*, Chris Callison-Burch, and Douglas Eck. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

Is this text machine-generated?

The University of Pennsylvania is the largest university in the United States and the second largest university in the world. Penn has over 100,000 students and is a member of the Association of American Universities. Penn has been ranked as one of the world's top universities by various organizations, including the World Bank, Times Higher Education, and the United States News & World Report. Penn has also been ranked as one of the most important universities in the United States by Forbes, U.S. News & World Report, and the Economist Intelligence Unit. Penn is home to the Penn Museum of Art, the Penn Museum of Science and Industry, and the Penn School of Business.

Is this text machine-generated? **Answer: yes**

The University of Pennsylvania is the largest university in the United States and the second largest university in the world. Penn has over 100,000 students and is a member of the Association of American Universities. Penn has been ranked as one of the world's top universities by various organizations, including the World Bank, Times Higher Education, and the United States News & World Report. Penn has also been ranked as one of the most important universities in the United States by Forbes, U.S. News & World Report, and the Economist Intelligence Unit. Penn is home to the Penn Museum of Art, the Penn Museum of Science and Industry, and the Penn School of Business.

Is this text machine-generated? **Answer: yes**

The University of Pennsylvania is the largest university in the United States and the second largest university in the world. Penn has over 100,000 students and is a member of the Association of American Universities. Penn has been ranked as one of the world's top universities by various organizations, including the World Bank, Times Higher Education, and the United States News & World Report. Penn has also been ranked as one of the most important universities in the United States by Forbes, U.S. News & World Report, and the Economist Intelligence Unit. Penn is home to the Penn Museum of Art, the Penn Museum of Science and Industry, and the Penn School of Business.

Is this text machine-generated? **Answer: yes**

The University of Pennsylvania is the largest university in the United States and the second largest university in the world. Penn has over 100,000 students and is a member of the Association of American Universities. Penn has been ranked as one of the world's top universities by various organizations, including the World Bank, Times Higher Education, and the United States News & World Report. Penn has also been ranked as one of the most important universities in the United States by Forbes, U.S. News & World Report, and the Economist Intelligence Unit. Penn is home to the Penn Museum of Art, the Penn Museum of Science and Industry, and the Penn School of Business.

Is this text machine-generated? Answer: yes

The University of Pennsylvania is the largest university in the United States and the second largest university in the world. Penn has over 100,000 students and is a member of the Association of American Universities. Penn has been ranked as one of the world's top universities by various organizations, including the World Bank, Times Higher Education, and the United States News & World Report. Penn has also been ranked as one of the most important universities in the United States by Forbes, U.S. News & World Report, and the Economist Intelligence Unit. Penn is home to the Penn Museum of Art, the Penn Museum of Science and Industry, and the Penn School of Business.

very unlikely ↔ very likely

What does the real Wikipedia article look like?

The University of Pennsylvania is a private Ivy League research university in Philadelphia, Pennsylvania. Established in 1740, it is the fourth oldest institution of higher education in the United States and among the highest ranked universities in the world. It is also one of nine colonial colleges chartered before the U.S. Declaration of Independence. Benjamin Franklin, the university's founder and first president, advocated for an educational institution that trained leaders in academia, commerce, and public service. Penn has four undergraduate schools as well as twelve graduate and professional schools.

very unlikely ↔ very likely

Do automatic classifiers and human raters detect the same text as machine-generated?

How does text generation work?

A **language model** predicts a probability distribution over what the next word in a sequence should be, given the previous words:

$$f_\theta(x_1, \dots, x_{t-1}) \sim P(X_t = v | x_1, \dots, x_t)$$

A **decoding method** is an algorithm for repeatedly selecting words to output given the language model's predicted distribution.

Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

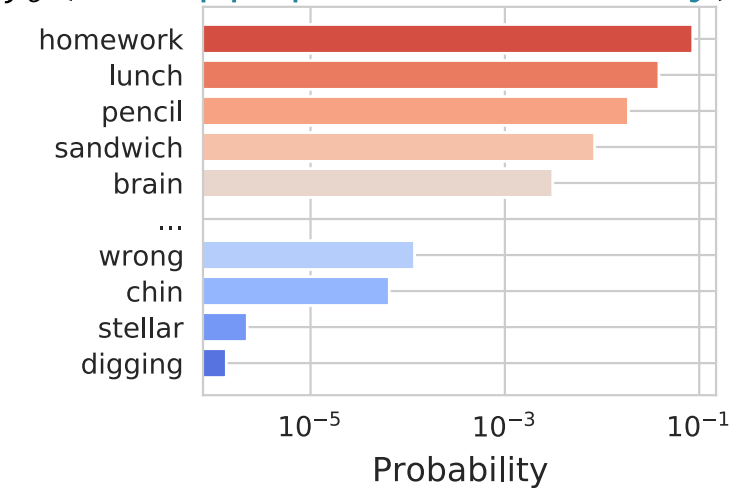
How does text generation work?

Suppose we want to generate a continuation for the prompt:

`A hippopotamus ate my`…

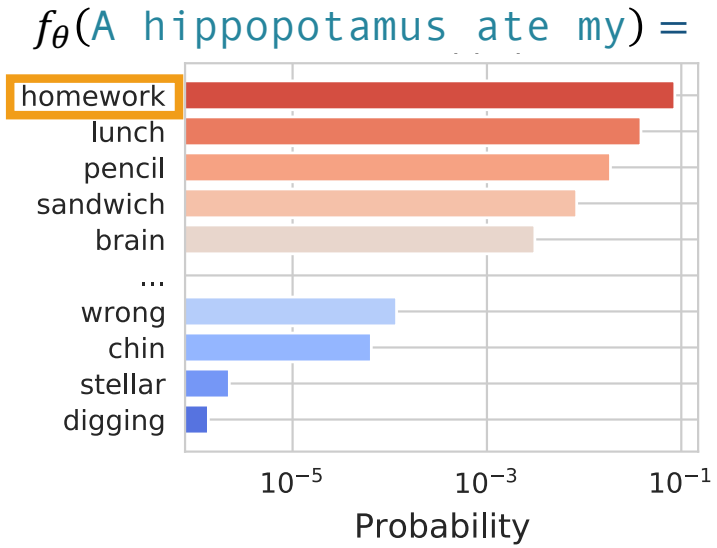
We compute:

f_θ(`A hippopotamus ate my`) =



Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

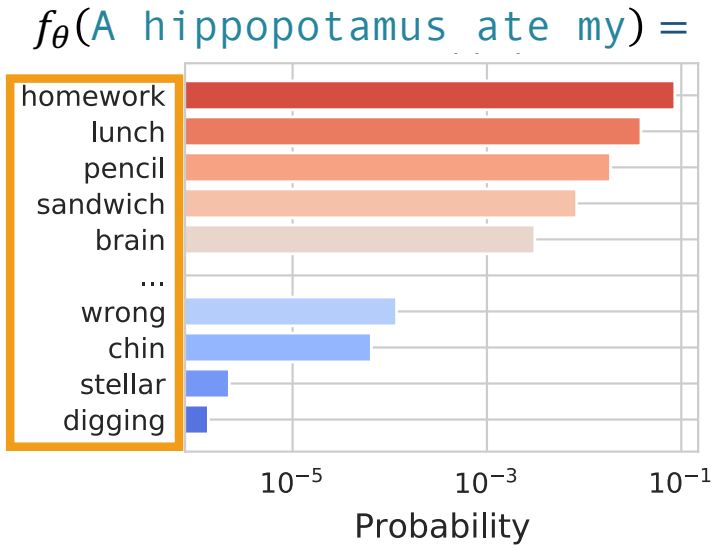
Three Common Decoding Methods



1. Always pick the most likely word.

Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

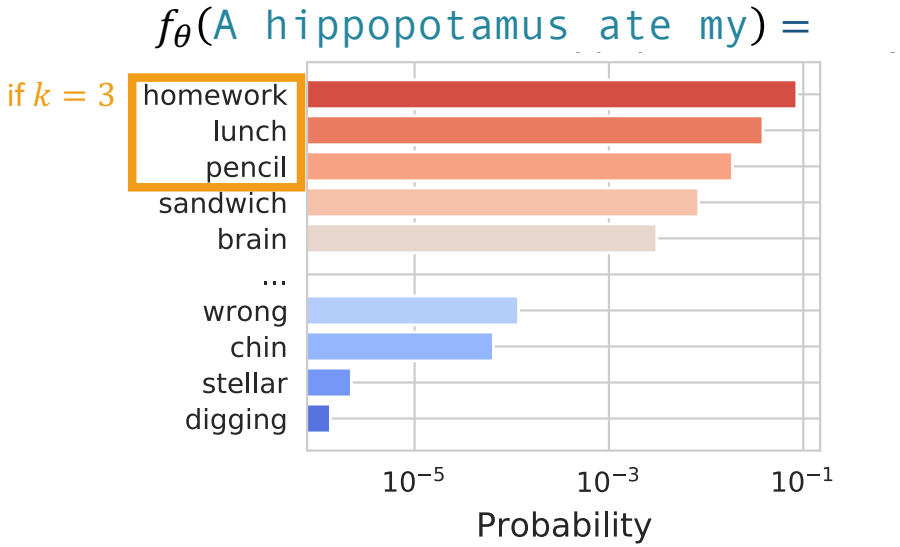
Three Common Decoding Methods



1. Always pick the most likely word.
2. Sample randomly from the full distribution predicted by the model.

Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

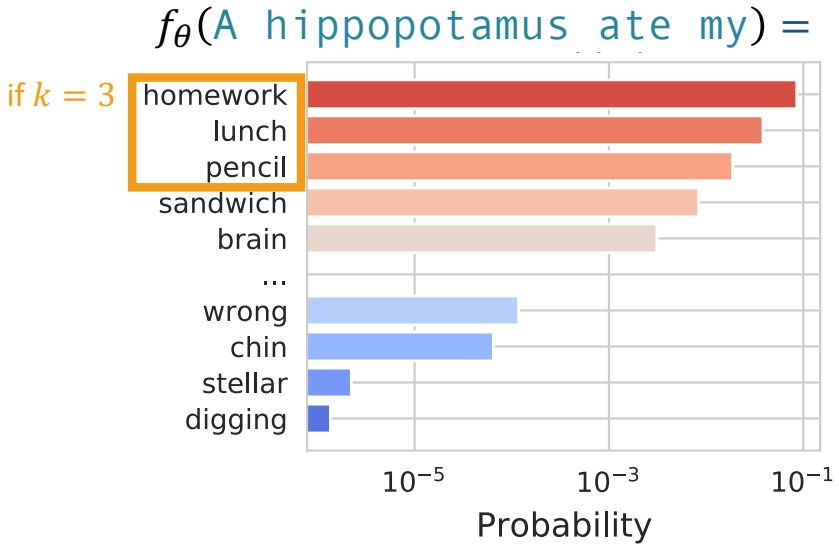
Three Common Decoding Methods



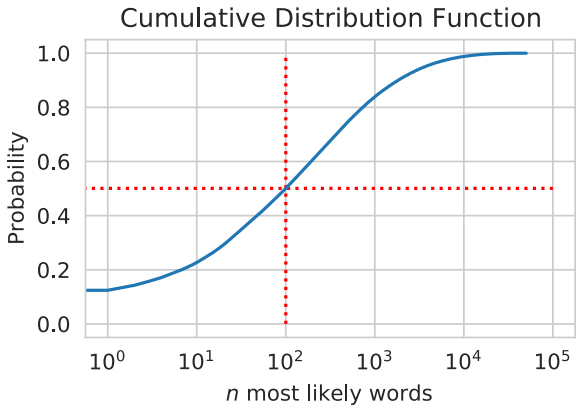
1. Always pick the most likely word.
2. Sample randomly from the full distribution predicted by the model.
3. Sample randomly from the k most likely words.

Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

Three Common Decoding Methods



1. Always pick the most likely word.
2. Sample randomly from the full distribution predicted by the model.
3. Sample randomly from the k most likely words.



Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

Three Common Decoding Methods

1. Always pick the most likely word.

A hippopotamus ate my homework. I was so embarrassed. I was so embarrassed. I was so embarrassed.

2. Sample randomly from the full distribution predicted by the model.

A hippopotamus ate my father's lunch. She was very amiable, shook her hands and smacked them together. Then she said: My, Seoudie, I was a hippopotamus too.

3. Sample randomly from the k=10 most likely words.

A hippopotamus ate my homework."

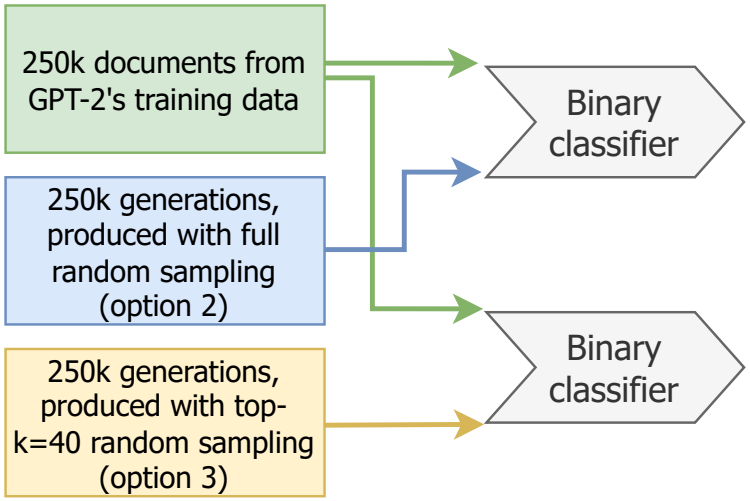
"A hippopotamus ate your homework?"

"Yes," said Jack. "I was in the library and I saw a hippo in the water. I thought it was a crocodile.

Daphne Ippolito. "A Tutorial on Neural Language Models and Text Generation." www.seas.upenn.edu/~daphnei/nlg_tutorial.pdf

Experimental Design

Automatic Detection

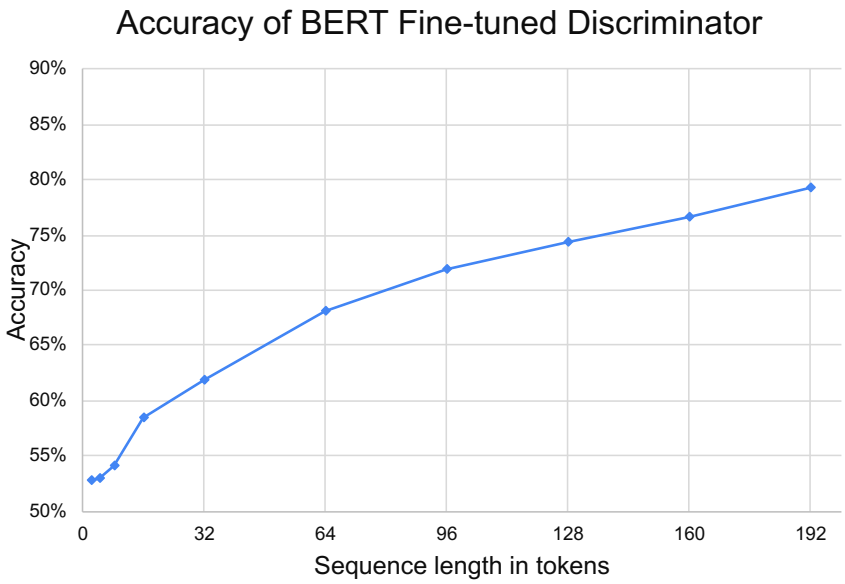


Human Detection

22 annotators were shown passages of text and asked to label them as human-written or machine-generated.

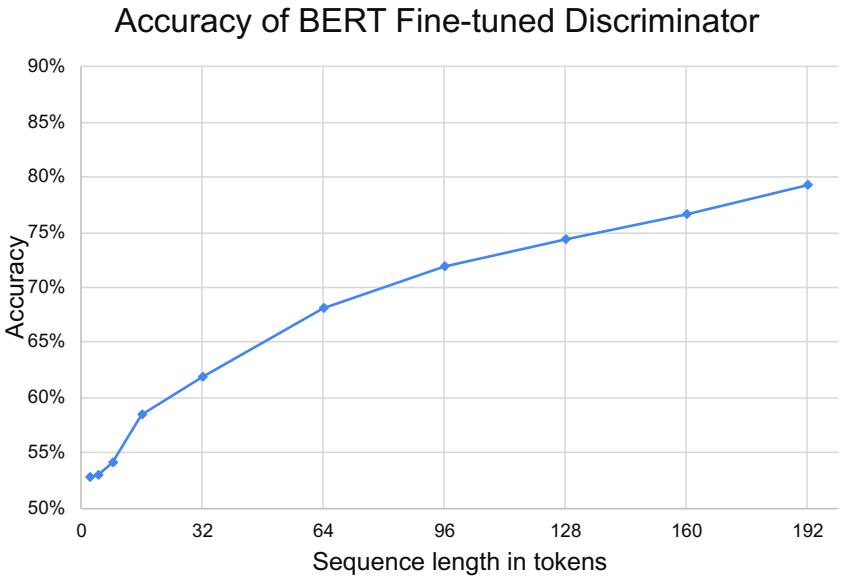
Daphne Ippolito*, Daniel Duckworth*, Chris Callison-Burch, and Douglas Eck. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

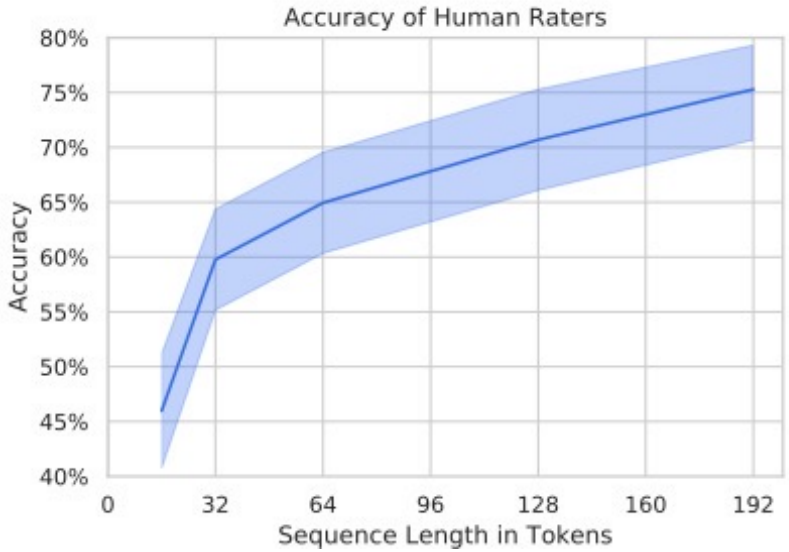
Results



Daphne Ippolito*, Daniel Duckworth*, Chris Callison-Burch, and Douglas Eck. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

Results

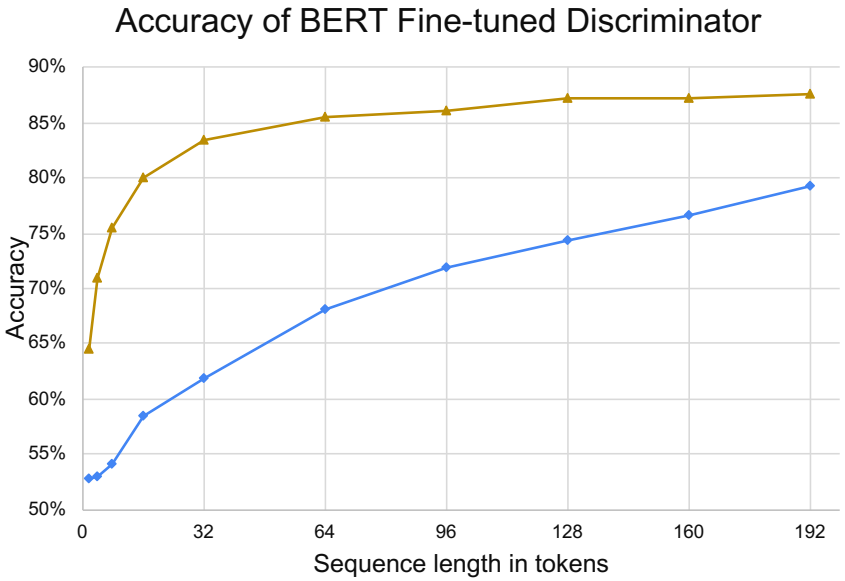


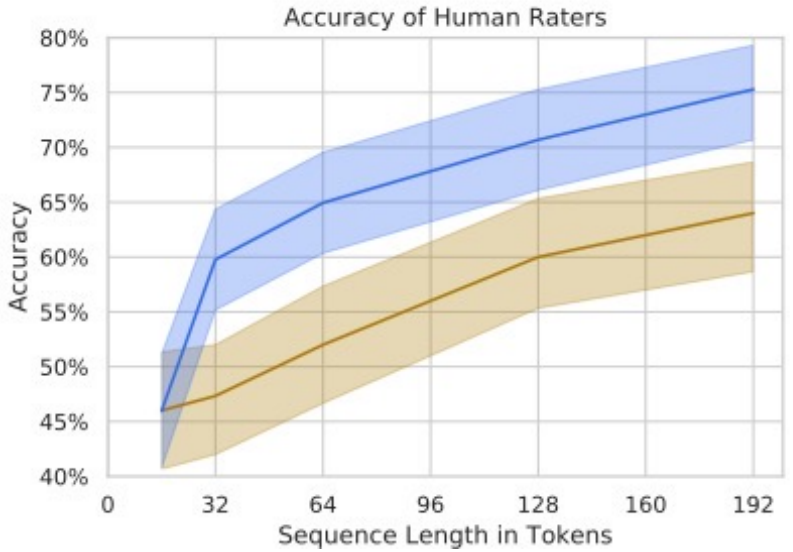


Sampling from full distribution

Daphne Ippolito*, Daniel Duckworth*, Chris Callison-Burch, and Douglas Eck. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

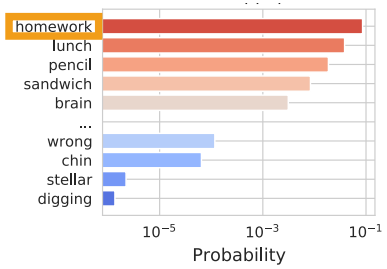
Results





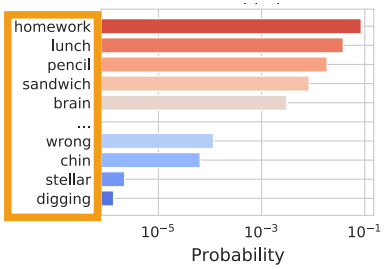
Sampling from the full distribution

Sampling from only the 40 most likely tokens at each step



$k = 1$

The Decoding Strategy Tradeoff



$k =$ vocab size

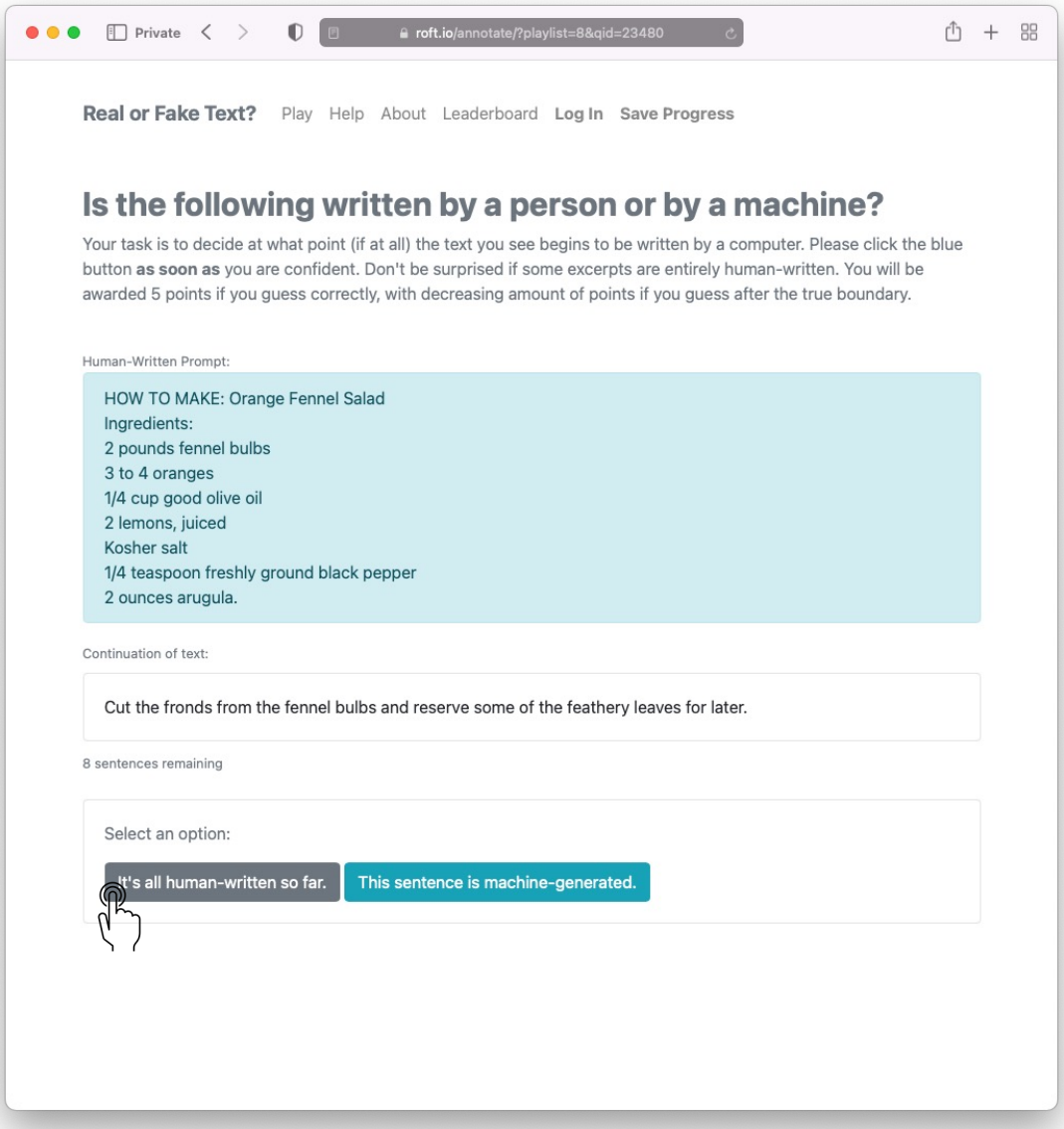
- Less diverse text.
- Fewer semantic errors.
- More likely to fool humans.
- Less likely to fool automatic classifiers.

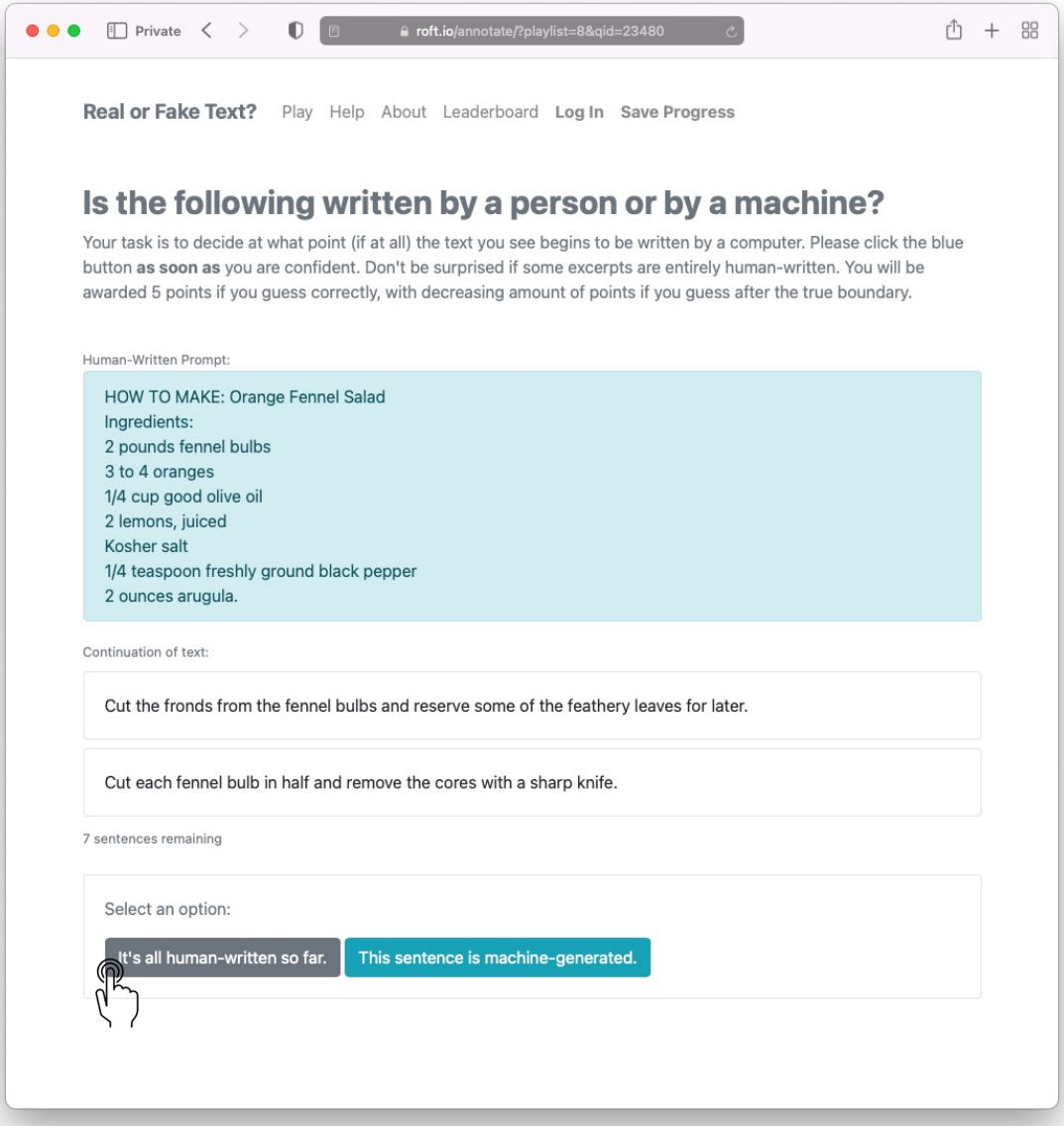
- More diverse text.
- More semantic errors.
- Less likely to fool humans.
- More likely to fool automatic classifiers.

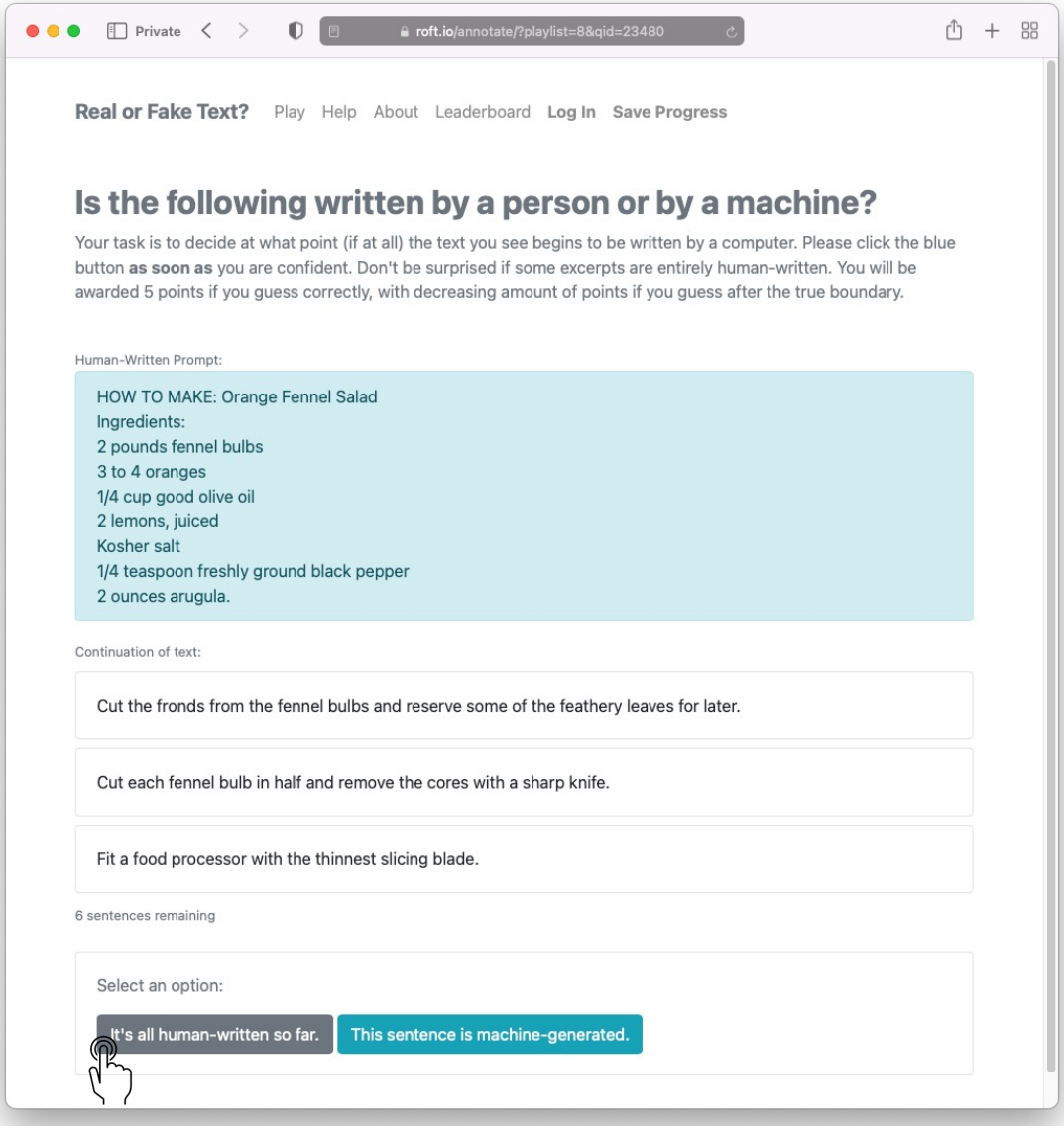
Caveats for Real World Automatic Detection

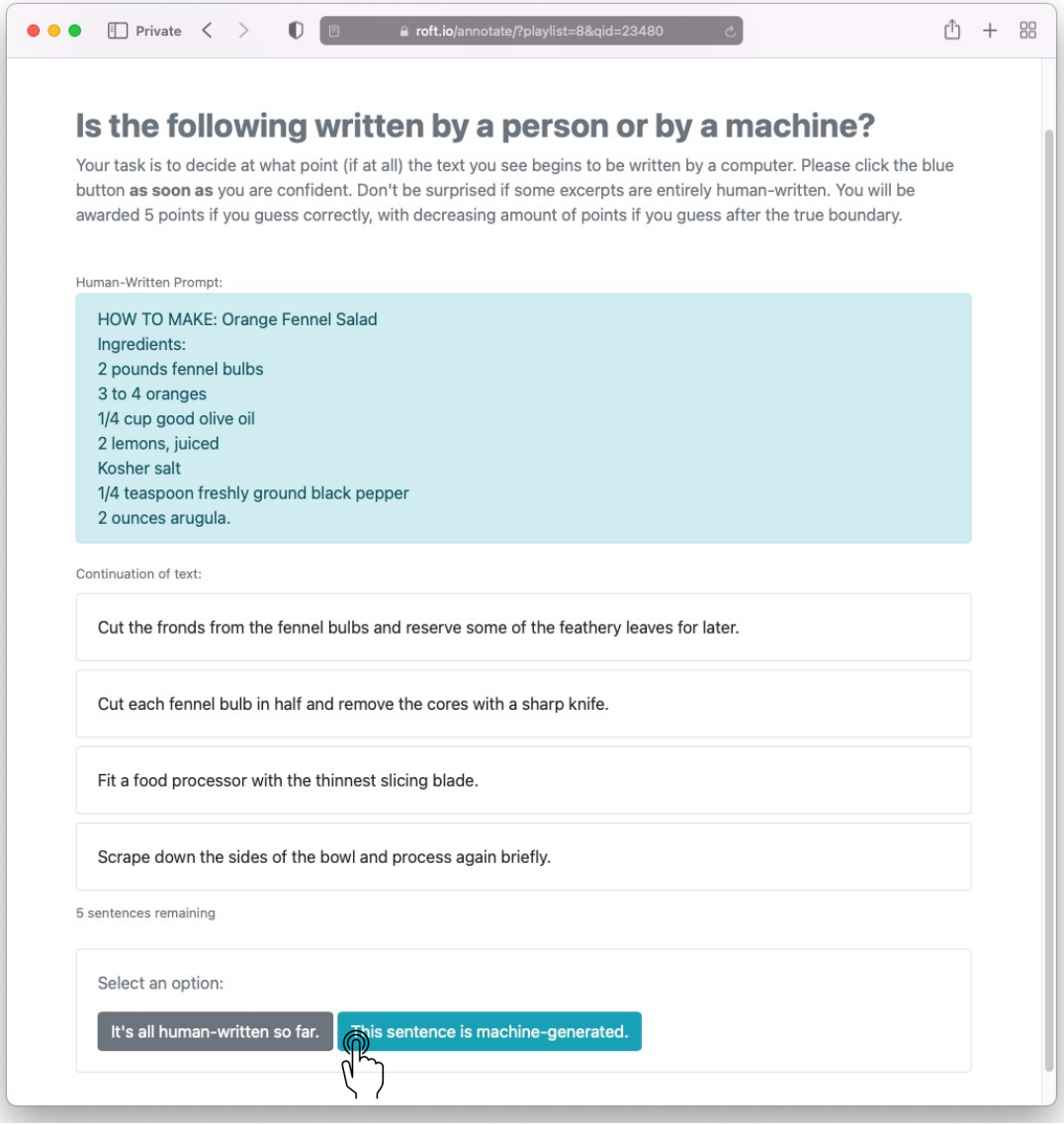
- Our work presents a best-case scenario.
  - What if the decoding strategy used for generation is unknown?
  - What if the model used for generation is unknown?
- Some domains are easier to detect generated text in than others.

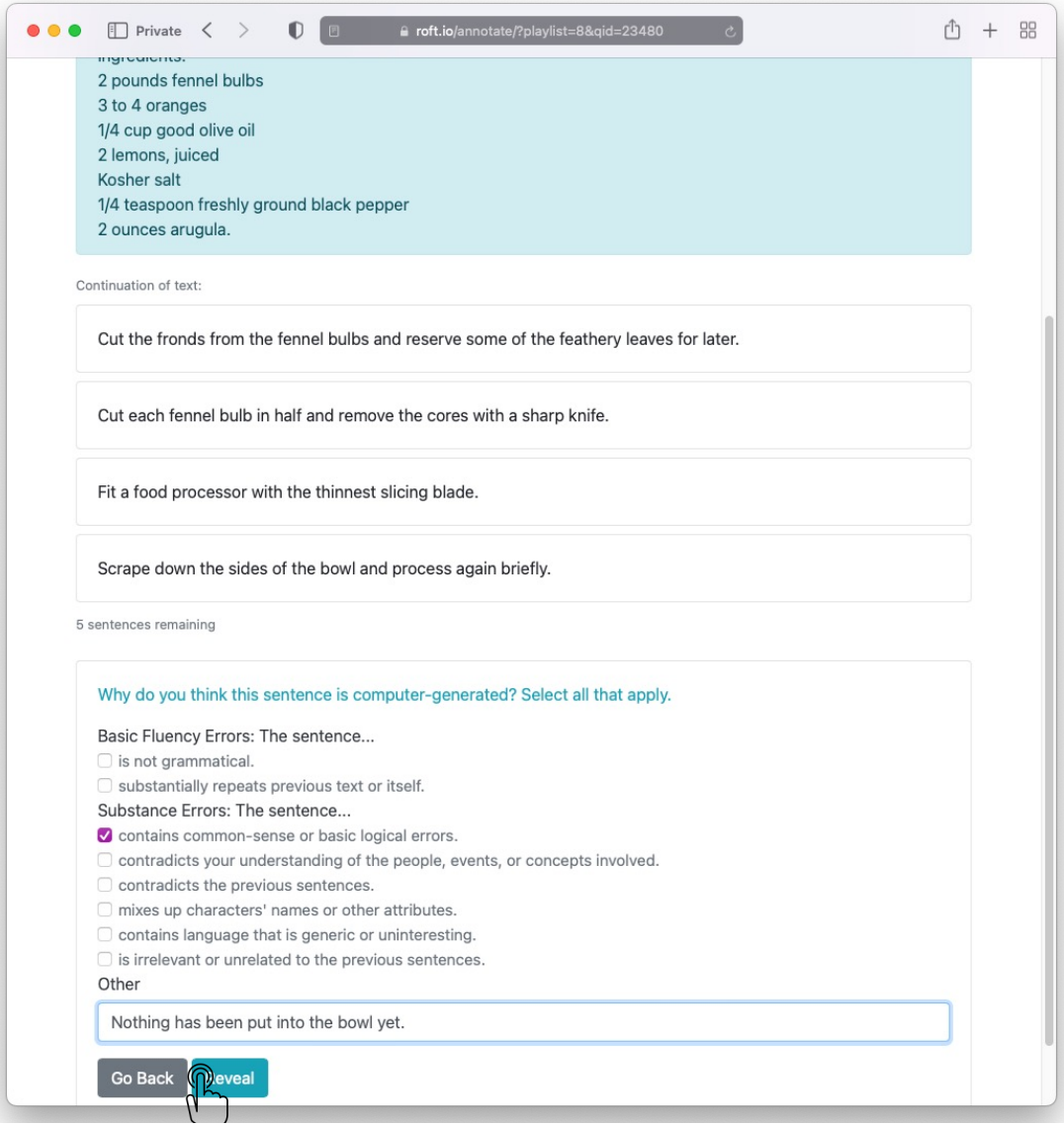
Can we gamify the detection text in order to conduct a largescale study of human detection ability?

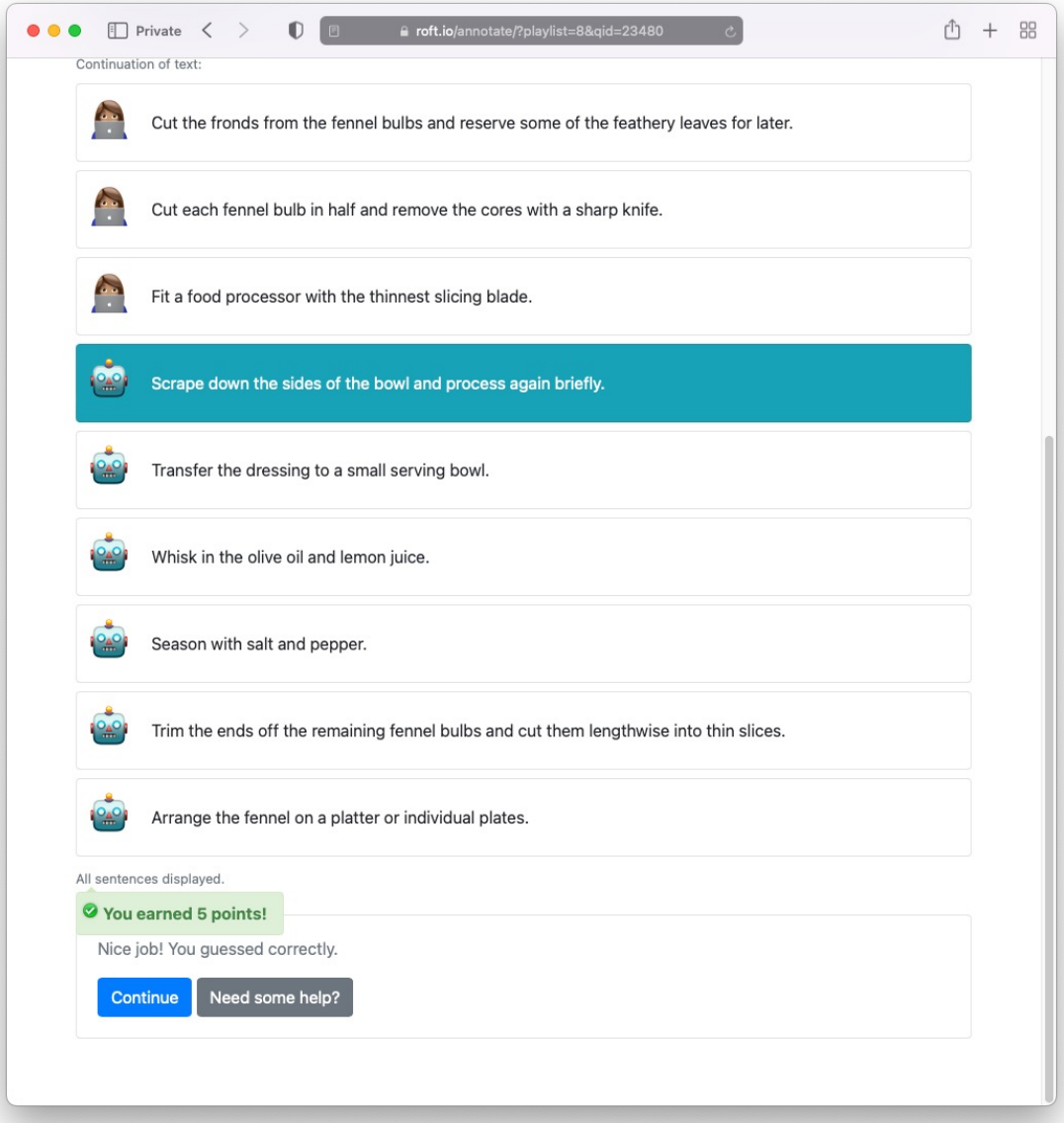

Private
roft.io/annotate/?playlist=8&qid=23480
Real or Fake Text?
Play
Help
About
Leaderboard
Log In
Save Progress
Is the following written by a person or by a machine?
Your task is to decide at what point (if at all) the text you see begins to be written by a computer. Please click the blue button as soon as you are confident. Don't be surprised if some excerpts are entirely human-written. You will be awarded 5 points if you guess correctly, with decreasing amount of points if you guess after the true boundary.
Human-Written Prompt:
HOW TO MAKE: Orange Fennel Salad
Ingredients:
2 pounds fennel bulbs
3 to 4 oranges
1/4 cup good olive oil
2 lemons, juiced
Kosher salt
1/4 teaspoon freshly ground black pepper
2 ounces arugula.
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
8 sentences remaining
Select an option:
It's all human-written so far.
This sentence is machine-generated.

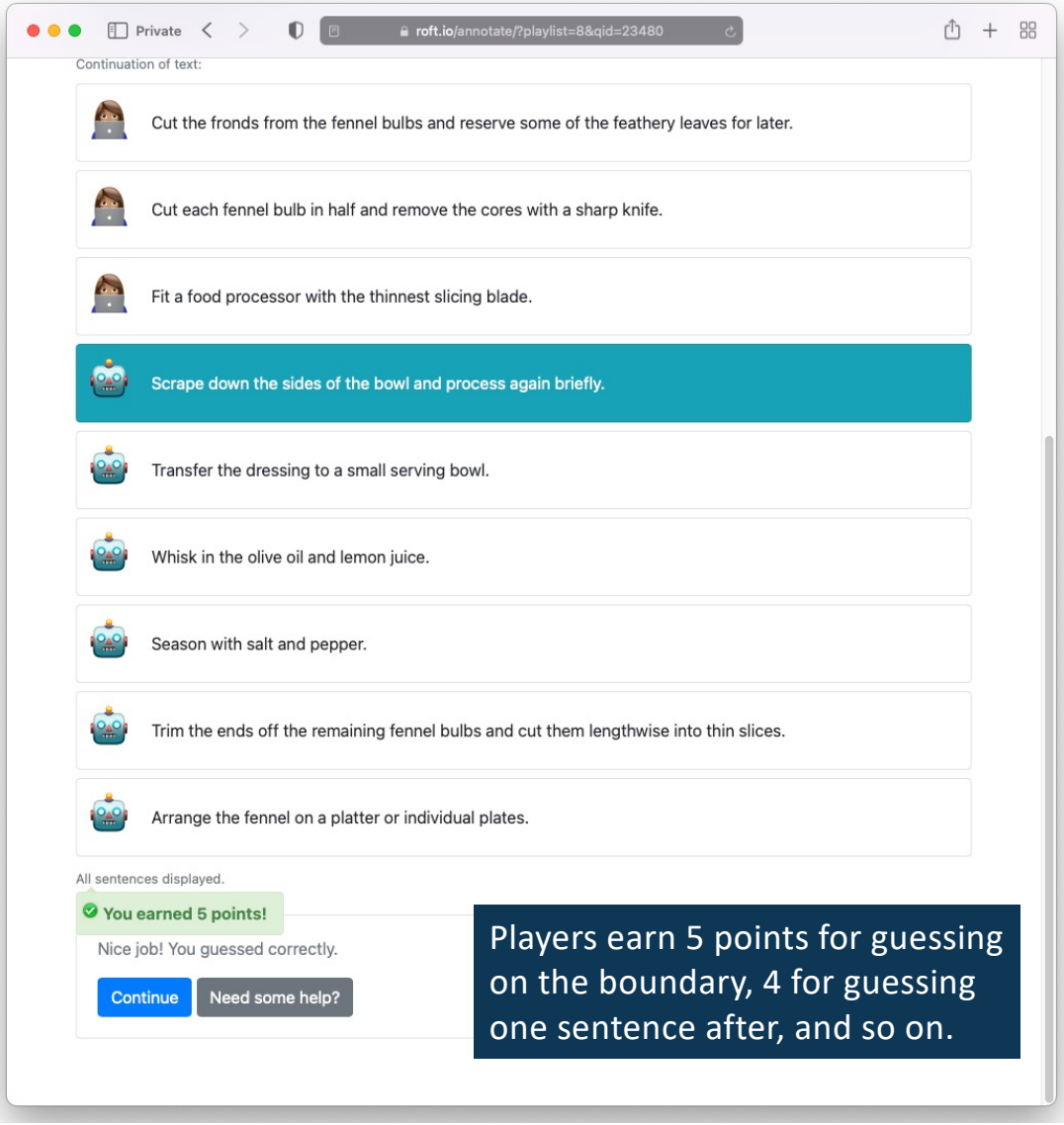

Private
roft.io/annotate/?playlist=8&qid=23480
Real or Fake Text?
Play
Help
About
Leaderboard
Log In
Save Progress
Is the following written by a person or by a machine?
Your task is to decide at what point (if at all) the text you see begins to be written by a computer. Please click the blue button as soon as you are confident. Don't be surprised if some excerpts are entirely human-written. You will be awarded 5 points if you guess correctly, with decreasing amount of points if you guess after the true boundary.
Human-Written Prompt:
HOW TO MAKE: Orange Fennel Salad
Ingredients:
2 pounds fennel bulbs
3 to 4 oranges
1/4 cup good olive oil
2 lemons, juiced
Kosher salt
1/4 teaspoon freshly ground black pepper
2 ounces arugula.
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
7 sentences remaining
Select an option:
It's all human-written so far.
This sentence is machine-generated.


Private
roft.io/annotate/?playlist=8&qid=23480
Real or Fake Text?
Play
Help
About
Leaderboard
Log In
Save Progress
Is the following written by a person or by a machine?
Your task is to decide at what point (if at all) the text you see begins to be written by a computer. Please click the blue button as soon as you are confident. Don't be surprised if some excerpts are entirely human-written. You will be awarded 5 points if you guess correctly, with decreasing amount of points if you guess after the true boundary.
Human-Written Prompt:
HOW TO MAKE: Orange Fennel Salad
Ingredients:
2 pounds fennel bulbs
3 to 4 oranges
1/4 cup good olive oil
2 lemons, juiced
Kosher salt
1/4 teaspoon freshly ground black pepper
2 ounces arugula.
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
Fit a food processor with the thinnest slicing blade.
6 sentences remaining
Select an option:
It's all human-written so far.
This sentence is machine-generated.


Private
roft.io/annotate/?playlist=8&qid=23480
Is the following written by a person or by a machine?
Your task is to decide at what point (if at all) the text you see begins to be written by a computer. Please click the blue button as soon as you are confident. Don't be surprised if some excerpts are entirely human-written. You will be awarded 5 points if you guess correctly, with decreasing amount of points if you guess after the true boundary.
Human-Written Prompt:
HOW TO MAKE: Orange Fennel Salad
Ingredients:
2 pounds fennel bulbs
3 to 4 oranges
1/4 cup good olive oil
2 lemons, juiced
Kosher salt
1/4 teaspoon freshly ground black pepper
2 ounces arugula.
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
Fit a food processor with the thinnest slicing blade.
Scrape down the sides of the bowl and process again briefly.
5 sentences remaining
Select an option:
It's all human-written so far.
This sentence is machine-generated.

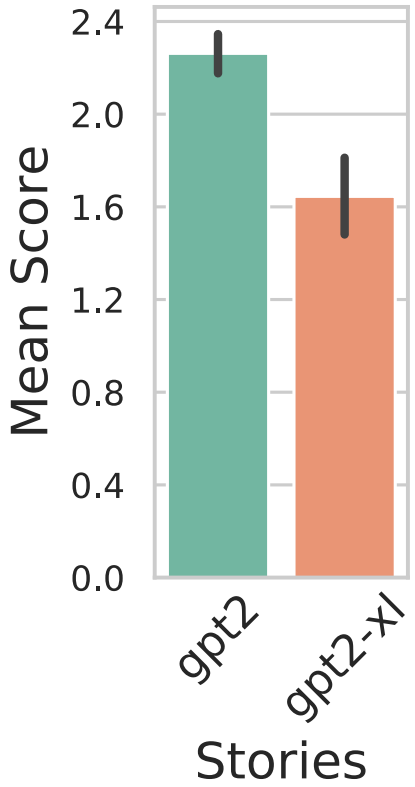

Private
roft.io/annotate/?playlist=8&qid=23480
2 pounds fennel bulbs
3 to 4 oranges
1/4 cup good olive oil
2 lemons, juiced
Kosher salt
1/4 teaspoon freshly ground black pepper
2 ounces arugula.
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
Fit a food processor with the thinnest slicing blade.
Scrape down the sides of the bowl and process again briefly.
5 sentences remaining
Why do you think this sentence is computer-generated? Select all that apply.
Basic Fluency Errors: The sentence...
is not grammatical.
substantially repeats previous text or itself.
Substance Errors: The sentence...
contains common-sense or basic logical errors.
contradicts your understanding of the people, events, or concepts involved.
contradicts the previous sentences.
mixes up characters' names or other attributes.
contains language that is generic or uninteresting.
is irrelevant or unrelated to the previous sentences.
Other
Nothing has been put into the bowl yet.
Go Back
eveal


Private
roft.io/annotate/?playlist=8&qid=23480
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
Fit a food processor with the thinnest slicing blade.
Scrape down the sides of the bowl and process again briefly.
Transfer the dressing to a small serving bowl.
Whisk in the olive oil and lemon juice.
Season with salt and pepper.
Trim the ends off the remaining fennel bulbs and cut them lengthwise into thin slices.
Arrange the fennel on a platter or individual plates.
All sentences displayed.
You earned 5 points!
Nice job! You guessed correctly.
Continue
Need some help?

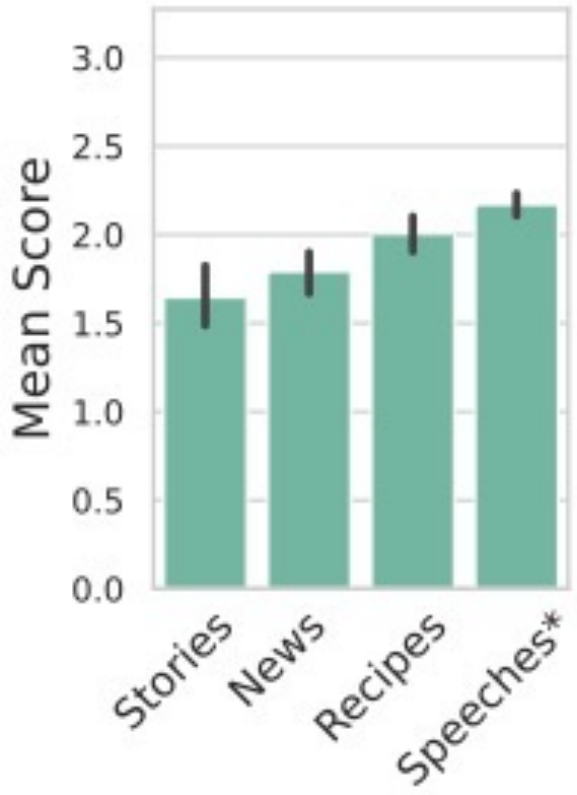

Private
roft.io/annotate/?playlist=8&qid=23480
Continuation of text:
Cut the fronds from the fennel bulbs and reserve some of the feathery leaves for later.
Cut each fennel bulb in half and remove the cores with a sharp knife.
Fit a food processor with the thinnest slicing blade.
Scrape down the sides of the bowl and process again briefly.
Transfer the dressing to a small serving bowl.
Whisk in the olive oil and lemon juice.
Season with salt and pepper.
Trim the ends off the remaining fennel bulbs and cut them lengthwise into thin slices.
Arrange the fennel on a platter or individual plates.
All sentences displayed.
You earned 5 points!
Nice job! You guessed correctly.
Continue
Need some help?
Players earn 5 points for guessing on the boundary, 4 for guessing one sentence after, and so on.

We collected over 21,000 annotations from 241 students taking an AI course at Penn.

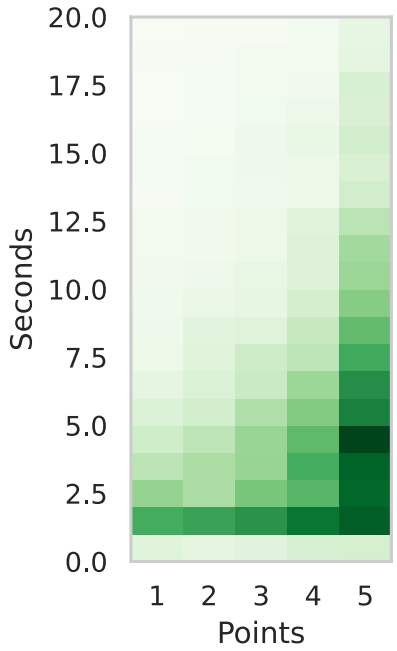
Results from RoFT Player Study



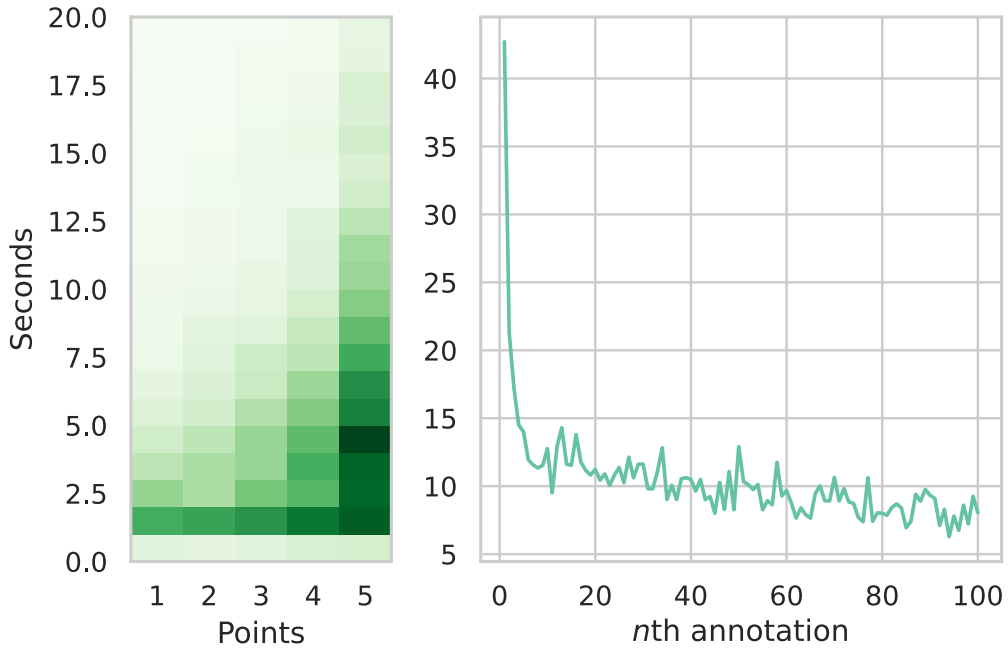
Results from RoFT Player Study



Results from RoFT Player Study



Results from RoFT Player Study



Is this text machine-generated?

Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

Is this text machine-generated? Yes and no.

Mr. and Mrs. Dursley of number four, Privet Drive, were proud to say that they were perfectly normal, thank you very much. They were the last people you'd expect to be involved in anything strange or mysterious, because they just didn't hold with such nonsense.

Mr. Dursley was the director of a firm called Grunnings, which made drills. He was a big, beefy man with hardly any neck, although he did have a very large mustache. Mrs. Dursley was thin and blonde and had nearly twice the usual amount of neck, which came in very useful as she spent so much of her time craning over garden fences, spying on the neighbors. The Dursleys had a small son called Dudley and in their opinion there was no finer boy anywhere.

Enabling Applications in Creative Writing

Detecting Machine-Generated Text

Measuring and Mitigating Memorization

Thesis Takeaways

When is memorization good vs. problematic?

- ✅ Outputting the home address of Joe Biden.
- ❌ Outputting the home address of Joe Smith.
- ✅ Correctly answering a question about the names of Harry Potter's aunt and uncle.
- ❌ Exactly generating the first chapter of Harry Potter.
- ✅ Exactly reproducing a well-known quote.
- ❌ Exactly reproducing a user's restaurant review.
- ✅ Providing accurate information on real businesses.
- ❌ Generating advertisements for real businesses.

In my dissertation work, I consider *all* generations which copy verbatim from the training set to be problematic.

Measuring Extractable Memorization

1. Select a text sequence from the train set, and divide it in two:

`It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness`

prefix | ground-truth continuation

2. Use language model to generate a continuation given the prefix.

3. The ground-truth continuation is considered *extractable* if it is exactly reproduced.

Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*.
"Quantifying Memorization Across Neural Language Models." In submission 2022.

Findings: Factors that Increase Memorization Tendencies

- Frequency of the sequence in the training set
  - More frequent sequences are more likely to be memorized.
- Size of the model
  - 6B param model memorizes as much as 3x more than 125M param model.
- Length of the prefix
  - Longer prefixes increase the ease of extracting the true continuation.

Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*. "Quantifying Memorization Across Neural Language Models." In submission 2022.

Three Strategies to Reduce Memorization

- Prevent memorized outputs from appearing during generation.

Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*.
"Quantifying Memorization Across Neural Language Models." In submission 2022.

Three Strategies to Reduce Memorization

- Prevent memorized outputs from appearing during generation.
- Use smaller models.
  - Models are rapidly getting bigger than their training data:
    - GPT-3 is a 175B parameter model. That's ~700 GB of model parameters.
    - GPT-3 was trained on 300 billion tokens of text. That's <600GB of training data.

Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*.
"Quantifying Memorization Across Neural Language Models." In submission 2022.

Three Strategies to Reduce Memorization

- Prevent memorized outputs from appearing during generation.
- Use smaller models.
  - Models are rapidly getting bigger than their training data:
    - GPT-3 is a 175B parameter model. That's ~700 GB of model parameters.
    - GPT-3 was trained on 300 billion tokens of text. That's <600GB of training data.
- Deduplicate training data.

Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*.
"Quantifying Memorization Across Neural Language Models." In submission 2022.

Training on deduplicated data reduces memorization.

(without hurting model quality!)

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Good Deduplication is Difficult

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Good Deduplication is Difficult

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Good Deduplication is Difficult

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
RealNews	KUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little louder lately as motorcycle makers, an aspiring middle class and easy bank credit come together to breed a new genus of motorcyclists – the big-bike rider. [...]	A visitor looks at a Triumph motorcycle on display at the Indonesian International Motor Show in Jakarta September 19, 2014. REUTERS/Darren Whiteside\nKUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little [...] big-bike rider. [...]

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Good Deduplication is Difficult

Dataset	Example	Near-Duplicate Example
Wiki-40B	\n_START_ARTICLE_\nHum Award for Most Impactful Character \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]	\n_START_ARTICLE_\nHum Award for Best Actor in a Negative Role \n_START_SECTION_\nWinners and nominees\n_START_PARAGRAPH_\nIn the list below, winners are listed first in the colored row, followed by the other nominees. [...]
LM1B	I left for California in 1979 and tracked Cleveland 's changes on trips back to visit my sisters .	I left for California in 1979 , and tracked Cleveland 's changes on trips back to visit my sisters .
RealNews	KUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little louder lately as motorcycle makers, an aspiring middle class and easy bank credit come together to breed a new genus of motorcyclists – the big-bike rider. [...]	A visitor looks at a Triumph motorcycle on display at the Indonesian International Motor Show in Jakarta September 19, 2014. REUTERS/Darren Whiteside\nKUALA LUMPUR (Reuters) - Roads in Southeast Asia have been getting a little [...] big-bike rider. [...]
C4	Affordable and convenient holiday flights take off from your departure country, "Canada". From May 2019 to October 2019, Condor flights to your dream destination will be roughly 6 a week! Book your Halifax (YHZ) - Basel (BSL) flight now, and look forward to your "Switzerland" destination!	Affordable and convenient holiday flights take off from your departure country, "USA". From April 2019 to October 2019, Condor flights to your dream destination will be roughly 7 a week! Book your Maui Kahului (OGG) - Dubrovnik (DBV) flight now, and look forward to your "Croatia" destination!

Previous Deduplication Efforts were Insufficient

- Chelba et al. (2013) Removed exact duplicate examples. [LM1B]
- Raffel et al. (2019) defined a “paragraph” as all text between two new lines. They removed paragraphs that were identical to any paragraph already in the dataset. [C4]
- Zellers et al. (2020) hashed the 50 characters of each example. Remove example with duplicate hashes. [GROVER]

Our Proposed Deduplication Methods

NEARDUP:

- Use an approximate matching algorithm, MinHash, to identify *examples* with high *n*-gram overlap.
- For each cluster of near-duplicate examples, we delete all but one.

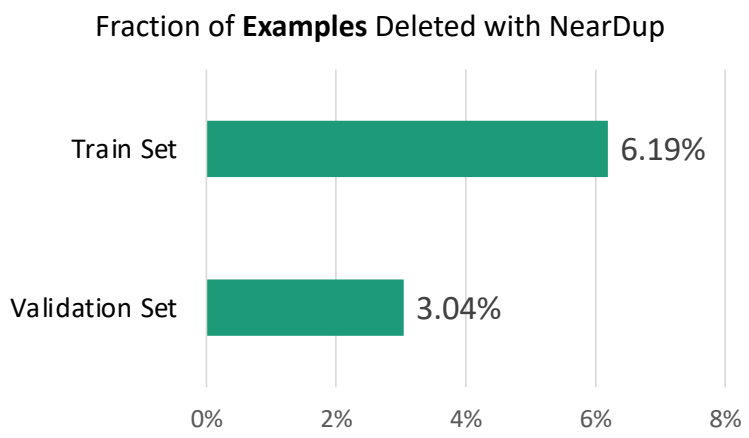
EXACTSUBSTR:

- Build a suffix array of the dataset to identify repeating *substrings* of at least 50 tokens.
- For each 50-token substring, we delete all but one instance.

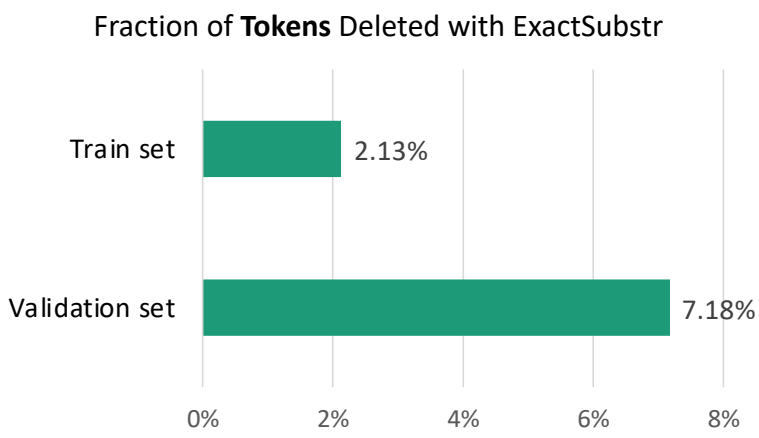
Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Our Proposed Deduplication Methods

NEARDUP: Approximate matching with MinHash



EXACTSUBSTR: Exact matching of document substrings using a suffix array



Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Experimental Design

- We trained three 1.5B parameter neural language models, one each on:
  - The original C4 dataset
  - C4 deduplicated with NEARDUP
  - C4 deduplicated with APPROXSBTR
- For each model, we measured:
  - The perplexity on evaluation datasets.
  - The amount of memorized content the model is capable of generating.

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Evaluating Language Models with Perplexity

- Perplexity is a measure of how likely the model thinks an evaluation dataset is.
- Normally in NLP, if a model achieves *lower* perplexity on a test or validation set, we say it is a *better* model.

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

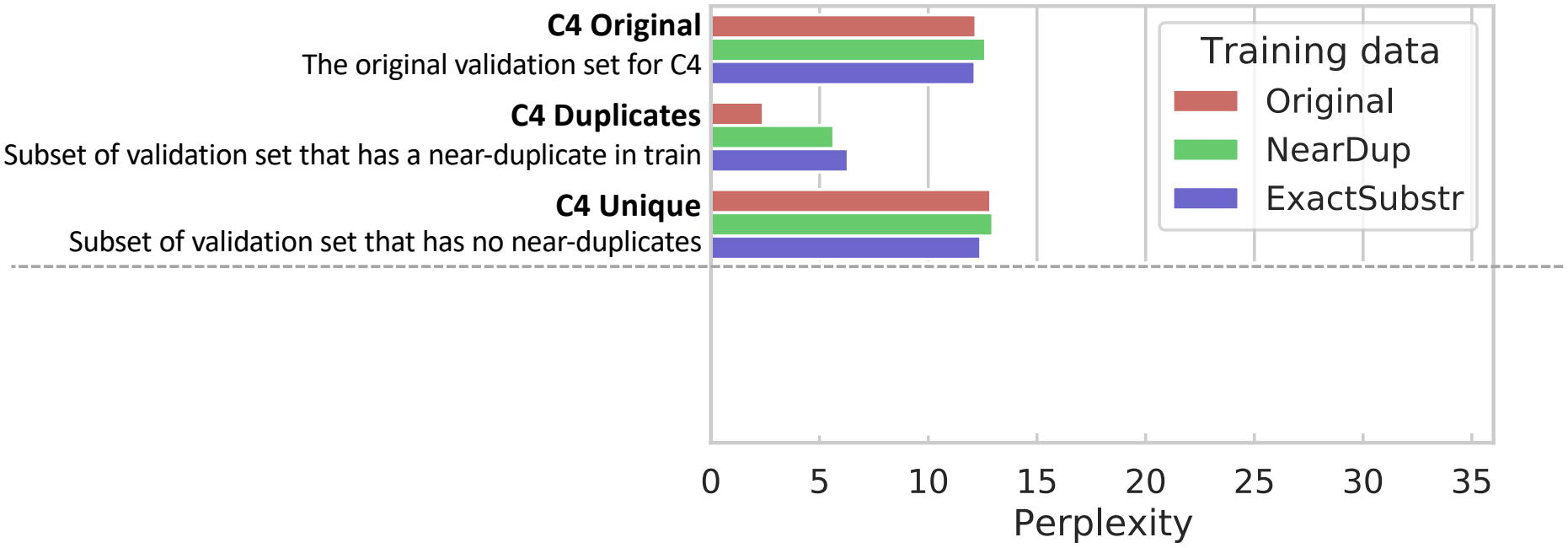
Evaluating Language Models with Perplexity

- Perplexity is a measure of how likely the model thinks an evaluation dataset is.
- Normally in NLP, if a model achieves *lower* perplexity on a test or validation set, we say it is a *better* model.

One of the most basic principles of machine learning is that one evaluates models on data that wasn't seen during training.

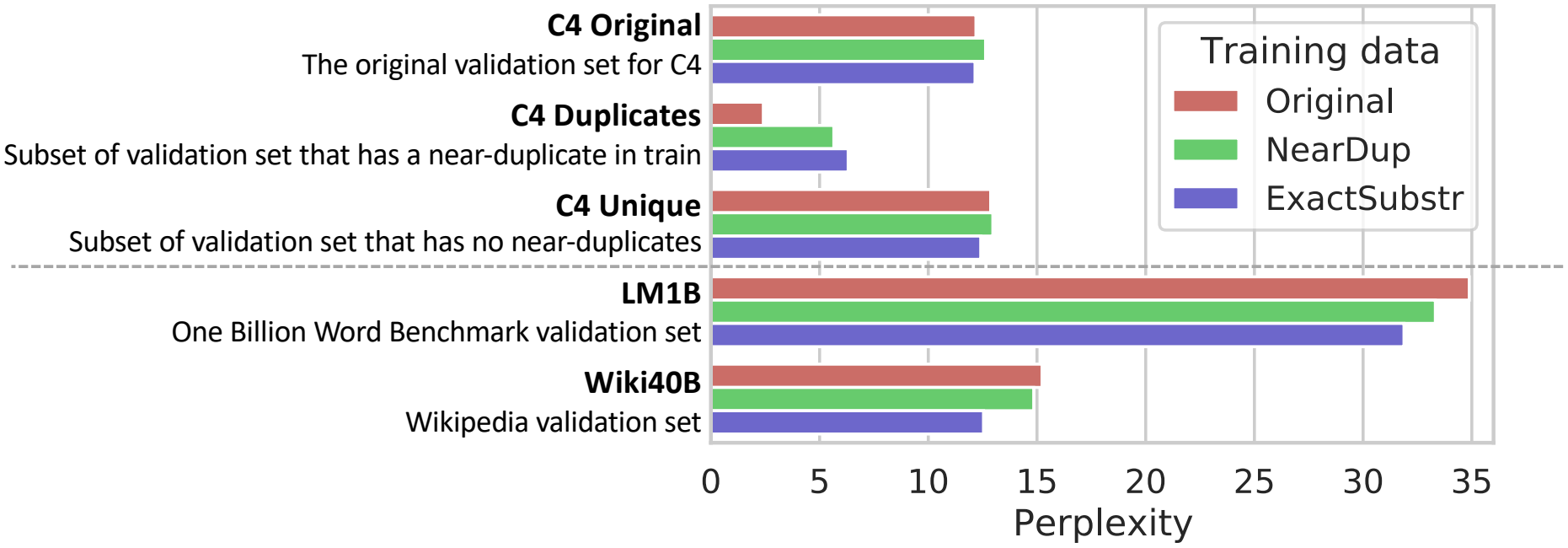
Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Impact of Deduplication on Perplexity



Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Impact of Deduplication on Perplexity



Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Impact of Deduplication on Unprompted Memorization

Training Data	% Generations memorized
Original C4	1.926%
C4 deduplicated with NEARDUP	0.189%
C4 deduplicated with EXACTSUBSTR	0.138%

By training on well-deduplicated data, we observed a 10x drop in instances of memorized content being generated.

Katherine Lee*, Daphne Ippolito*, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. "Deduplicating Training Data makes Language Models Better." In ACL 2022.

Language Generation and Creativity
Detecting Generated Text
The Risks of Memorization
Thesis Takeaways

Text generation involves tradeoffs.

- We discussed the quality-diversity tradeoff.
- Other tradeoffs:
  - Language that is safe vs. language that is disruptive
  - Factual vs. fantastical
  - Multiple specialized language models vs. one general-purpose one
  - Smaller vs. larger models
  - Smaller but higher-quality training set vs. larger but lower-quality one

More attention needs to be paid to training dataset selection.

- Much of the training data for language models is derived from scraping the web.
- Duplicate text is only one problem in these datasets.
- Other challenges include:
  - How stringent should toxicity filters be?
  - What languages should be included and how do we source text in them?
  - How should we balance different genres/sources (news, Wikipedia, books, etc.)?

Detection of generated text is getting harder but no less important.

- Why harder:
  - In my work on automatic detection, we took advantage of the fact that convincing generated text has more high-likelihood words in it than what a human would write.
  - This bias will decrease as models get better.
- Why important:
  - There have been real-world instances of bad actors passing off generated text as genuine.
  - Generated text can sully future training sets.

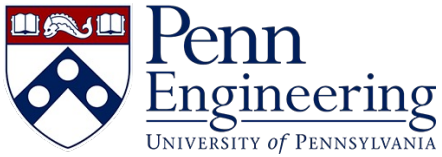
Memorization and plagiarism are serious concerns but hard to define.

- Given two text sequences, us humans can make judgement calls as to whether they are the same.
- Existing methods to measure document similarity are imperfect proxies of this human judgement.
- For example, are these the same?
  - `The Hippopotamus and the Kitten Traveled through the Milky Way.`
  - `The hippopotamus & the kitten traveled thru the milky way`

Evaluation in real-world settings reveals the weaknesses of state-of-the-art NLG.

- Language models need to improve at
  - Understanding long text passages.
  - Controllable style and voice.
  - Opinionated and risky generation.
- User interface is critical.

Thank you to all my collaborators.



Chris-Callison Burch
Liam Dugan
Arun Kirubarajan
Reno Kriz
Maria Kustikova
Joao Sedoc

Douglas Eck
Nicholas Carlini
Andy Coenen
Daniel Duckworth
David Grangier
Katherine Lee

Emily Reif
Matthew Jagielski
Florian Tramèr
Ann Yuan
Chiyuan Zhang

Questions?

Expanding NLG's Capabilities for Positive Impact

Ann Yuan, Andy Coenen, Emily Reif, Daphne Ippolito. "Wordcraft: Story Writing with Large Language Models." In IUI 2022.

Daphne Ippolito, et al. "A Recipe for Arbitrary Text Style Transfer with Large Language Models." In *ACL* 2022.

Daphne Ippolito, et al. "The Case for a Single Model that can Both Generation Continuations and Fill in the Blank." In NAACL 2022.

Daphne Ippolito, et al. "Toward Better Storylines with Sentence-Level Language Models." In *ACL* 2020.

Daphne Ippolito, et al. "Unsupervised Hierarchical Story Infilling" In *ACL Workshop on Narrative Understanding* 2019.

Privacy Considerations of Digital Contact Tracing

Hyunghoon Cho, Daphne Ippolito, Yun William Yu. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. Preprint arXiv:2003.11511. 2020.

Yoshua Bengio, Richard Janda, Yun William Yu, Daphne Ippolito, et al. "The need for privacy with public digital contact tracing during the COVID-19 pandemic." In *Lancet Digital Health* 2020.

Yoshua Bengio, Daphne Ippolito, et al. Inherent privacy limitations of decentralized contact tracing apps. In *JAMIA* 2021.

What makes deep fakes convincing?

Daphne Ippolito, et al. "Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text." In submission 2022.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, et. al., "Trading Off Diversity and Quality in Natural Language Generation." In *EACL HumEval Workshop* 2021.

Daphne Ippolito, et al. "RoFT: A Tool for Evaluating Human Detection of Machine-Generated Text." In *EMNLP System Demonstrations* 2020.

Daphne Ippolito, et al. "Automatic Detection of Generated Text is Easiest when Humans are Fooled." In *ACL* 2020.

Daphne Ippolito, et. al. "Comparison of Diverse Decoding Methods from Conditional Language Models." In *ACL* 2019.

Measuring and Mitigating the Risk of Memorization

Daphne Ippolito, et. al. "Deduplicating Training Data Makes Language Models Better." In *ACL* 2022.

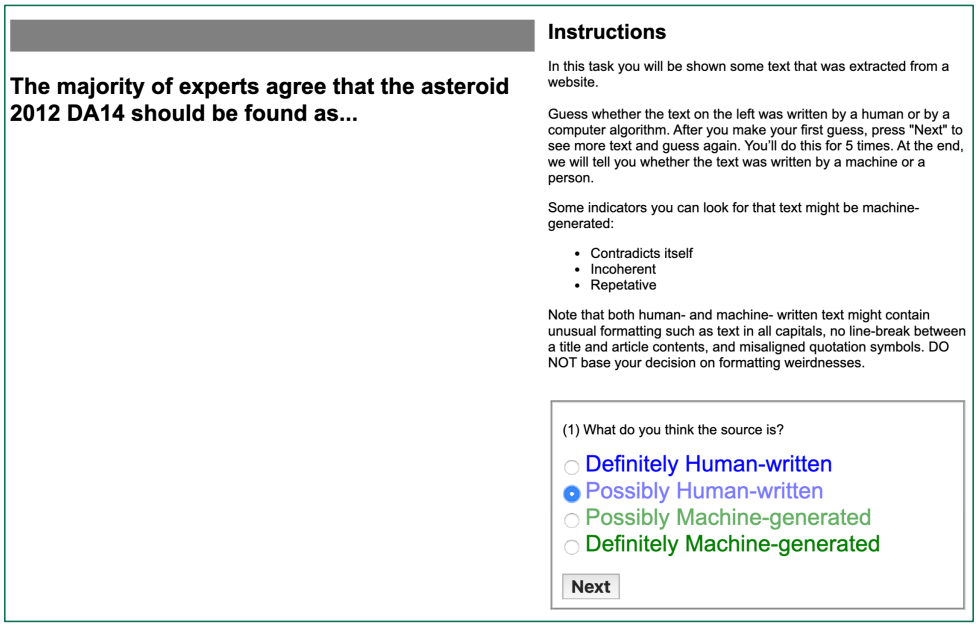
Chiyuan Zhang, Daphne Ippolito, et al. "Counterfactual Memorization in Neural Language Models." In submission 2022.

Nicholas Carlini, Daphne Ippolito, et. al. "Quantifying Memorization across Neural Language Models." In submission 2022.

EXTRA SLIDES

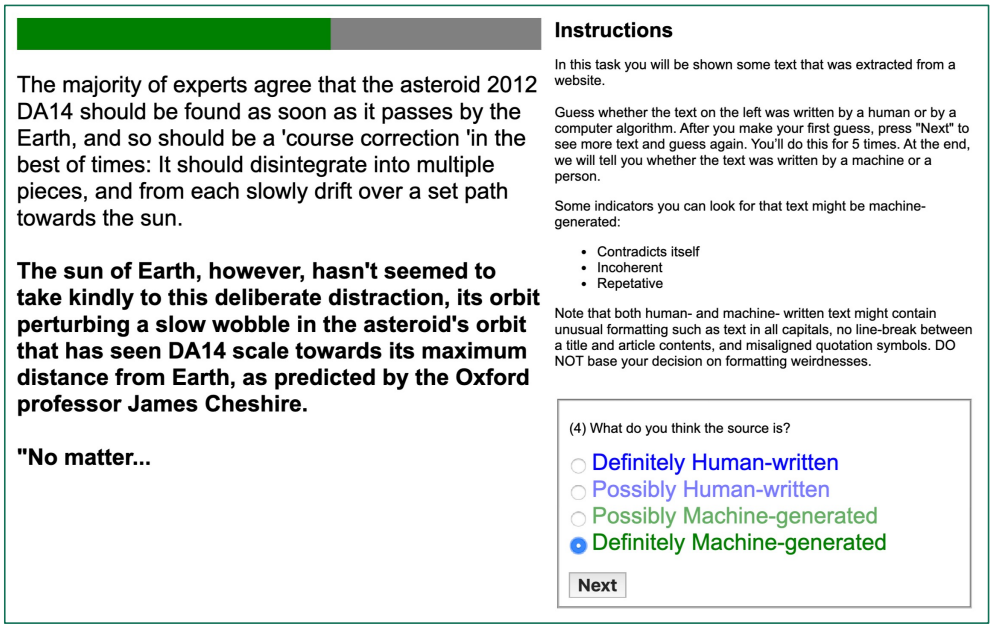
Experimental Design for Human Detection

- 22 raters labeled >500 excerpts.

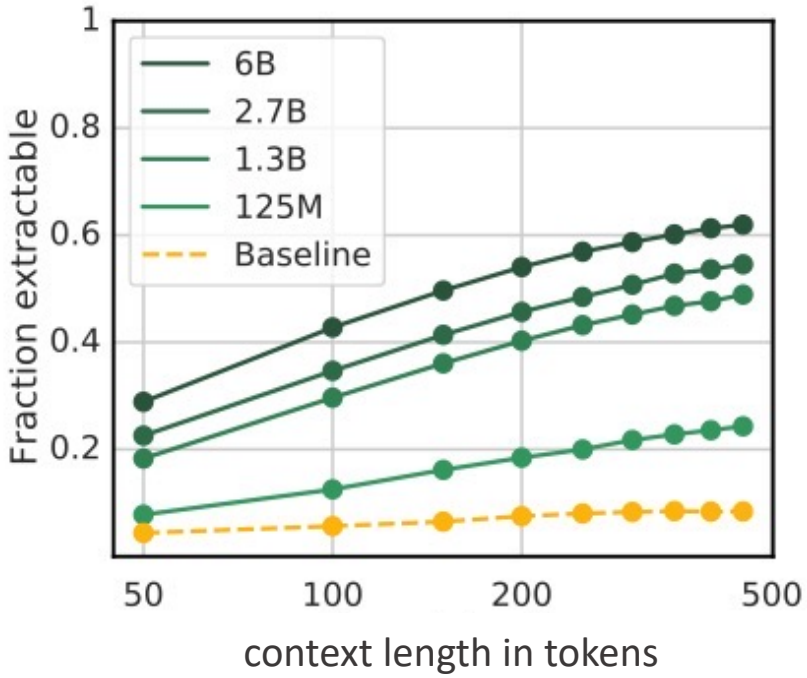


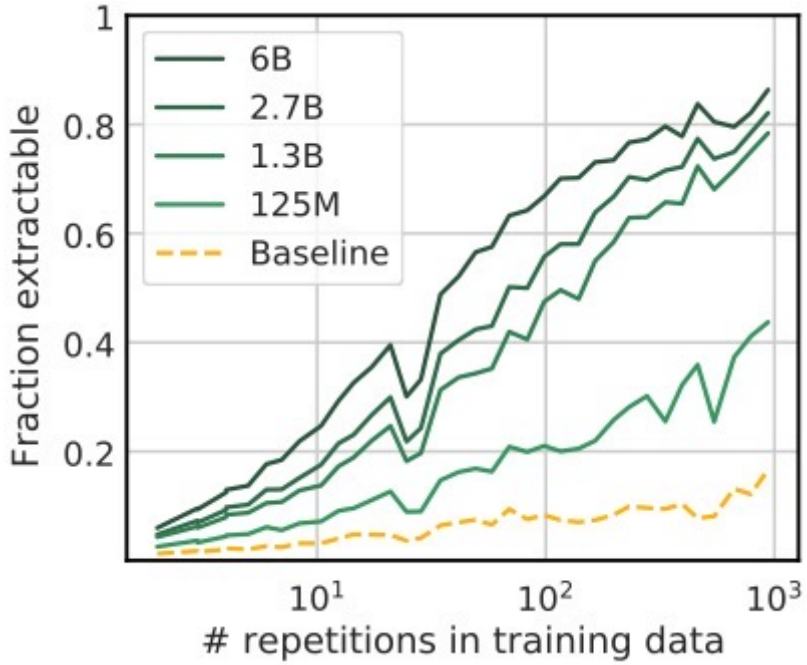
Experimental Design for Human Detection

- 22 raters labeled >500 excerpts.



Factors that increase likelihood of memorization

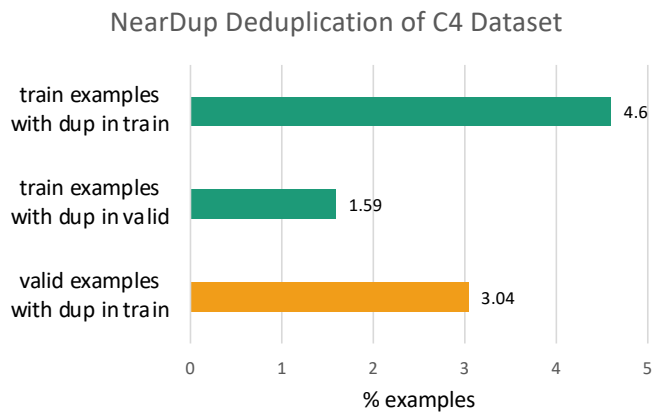




Nicholas Carlini*, Daphne Ippolito*, Matthew Jagielski*, Katherine Lee*, Florian Tramer*, Chiyuan Zhang*. "Quantifying Memorization Across Neural Language Models." In submission 2022.

Our Proposed Deduplication Methods

NEARDUP: Approximate matching with MinHash



EXACTSUBSTR: Exact matching of document substrings using a suffix array