Real or Fake Text Guide

The goal of RoFT is to conduct research into how hard it is to detect when a text switches from human writing to being written by an AI. This guide is designed to give you a sense of the kind of things you should look out for when trying to identify whether a sentence was machine-generated.

How does RoFT work?

How a game round works

1. In each game round you are shown the first sentence of a real human-written text passage.
2. You will be shown subsequent sentences one at a time. For each one, make a guess at whether this is a continuation of the real human-written passage or else a sentence was written by an AI. Note that some passages will be entirely human-written.
3. After you guess AI, pick a reason (or multiple) you think the text was machine-generated.
4. You will then be shown the true authorship of each sentence. Try to use this to learn to do better for next time.

How scoring works

In each game round, you will be awarded point in the following way:

- You will get 5 points for correctly guessing at exactly the boundary sentence--that is the first sentence of text generated by an AI.
- You will earn 5-k points if you guess k sentences after the true boundary.
- You will earn 0 points for guessing a human-written sentence is machine-generated.

You should do your best to guess “machine-generated” at the first sentence where you are fairly confident that an AI generated the text.

Signs of machine-generated text

Here are some of the signs you can look for to tell that a sentence might be machine-generated. Note that none of these are hard rules; rather they are general trends we have observed in machine-generated text.

[1] Text that is not grammatical.
Sometimes generated text will have weird, grammatical disfluencies. But remember, humans can write ungrammatically too (especially when they're posting on the internet).

[2] Text that substantially repeats previous text or itself.
Sometimes the AI system gets stuck in loops and writes the same thing multiple times in a row, or it repeats the same phrase multiple times but in slightly different ways.

[3] Text that is irrelevant or unrelated to the previous sentences.
Often, machine-generated text will meander into a new topic and never return to the original topic.

[4] Text that contradicts the previous sentences.
Contradictions are one of the easiest ways to detect machine-generated text. For example, a generated recipe instruction will list an ingredient which wasn't in the ingredients list, or a sentence will be generated that says the exact opposite of a previous sentence.

[5] Text contradicts your understanding of the people, events, or concepts involved.
If text mentions a real person, place, event, or other concept, but gets the details about it wrong, this could mean it is machine-generated.

[6] Text that contains common-sense or basic logical errors.
There are errors that violate your basic understanding of how the world works. For example, a generation about a unicorn with 4 horns or the sun brightly shining in the middle of the night.

[7] Text that mixes up characters' names or other attributes.
Especially, in the fiction domain, when there are multiple characters, generated text will tend to mix them up, for example, swapping characters' genders or physical traits.

[8] Text that contains language that is generic or uninteresting.
This category of errors is pretty subtle. Generation systems will tend to produce text that uses a lot of words to say not very much, since they are "rewarded" for producing common, simplistic language.

Features that may be misleading

Incorrect sentence boundary segmentation

Our game tries to show you one sentence at a time. Sentence boundaries were automatically parsed, and could be wrong. Errors in sentence segmenting do not imply the text was machine-generated.

Bad writing
In domains like New York Times articles, bad writing may indicate that the text was machine-generated, but in domains like the Reddit Short Stories, it may not. Remember that amateur human writers often write text that is incoherent, typo-ridden, and otherwise not very good.

Some concrete examples

Red is the human-written text and blue is the machine-generated continuation.

Example 1
Michigan 16, Ohio State 13 COLUMBUS, Ohio (AP) -- J. D. Carlson kicked a 37-yard field goal with three seconds left after Michigan had stopped Ohio State on a fourth-and-one situation with less than two minutes left, giving the Wolverines a Big Ten conference victory. Two plays later, a 17-yard interception return by Brian Poynter put the Buckeyes back at midfield. That set up one final play that would have made it 20-17, but John Tressel's 4-yard pass to Kenyatta Baldwin gave Michigan possession again with 8:53 remaining.

This example both contradicts previous sentences ([4]) and contradicts our understanding of the people and places involved ([5]). In this passage, the last human written sentence ends with “giving the Wolverines a Big Ten conference victory” but the following machine generated sentence implies that the game is still going on. While it is plausible (but unlikely) that Michigan could have immediately thrown an interception after regaining possession, the next generated sentence misrepresents both the score (20-17) and the time remaining (8:53).

Example 2
Roland's of Lebanon is hardly prepossessing. Its dimly lighted, exceedingly modest storefront locale is as bare bones as a sit-down dining room can be -- just a handful of tables, seating at most 32, with a counter area at the entrance where pizza can be ordered to take out. The inside sports neon signs that have the whole establishment set off, some of them directing patrons toward each others' stations, but both sides are full and it looks very cramped; certainly you won't want any friends staying after for drinks.

This example contains subtle incoherence which most closely falls into [6] (Text contains common-sense or basic logical errors). It has the following common-sense issue: why would there be neon lights directing patrons toward *each other’s* stations? This is not how we understand restaurants to work.

Example 3

HOW TO MAKE: Best Ever Potato Casserole Cullen Baker Country Sampler
Ingredients:
32 ounces frozen hash browns, thawed
1 (8 ounce) can cream of chicken soup
12 ounces Velveeta cheese
8 ounces sour cream
12 cup butter, melted
1 teaspoon salt
1 medium onion
2 cups corn flakes.
Place hash browns in 13*9 baking dish. Mix Soup, cheese, sour cream, salt, and onion together. I've never been able to get past the idea that they're just mushy peas with a few chunks of meat in them. So when I heard about this recipe from my friend, Lisa, I was skeptical. But after trying it, I realized how much better it is than anything else I've ever had. It's so creamy and full of flavor! The only thing missing is the crispy bits on top. That would be awesome too! And since we were making it ahead of time, we didn't have to worry about those pesky potatoes getting soggy. Plus, I know you don't want your potatoes to get all soft and mushy in the middle anyway! This is one of those recipes where you can make as many as you like.

This example contains text that is irrelevant or unrelated ([3]), contradictory ([4]), and generic or uninteresting ([8]). Despite the recipe not containing any peas or “chunks of meat” the very first generated sentence references these ingredients. In addition, the text rambles on about irrelevant aspects of the recipe without actually returning to the steps of the recipe.

It is worth noting that a digression to first person narrative is actually quite common in human-written recipes and should not be used as an indication of text being machine-generated. However, in this example, the text never returns to talking about the recipe, and thus is clearly generated.

Example 4

I had never used my powers before, in all honesty I was scared of what could happen. You know that whole butterfly effect. I knew my mother had been a time traveler, but I didn't know she passed the ability on to me, I suppose she never got the chance., I suppose she never got the chance. If she did we wouldn't be here at this exact moment. Taylor nodded. "So as much as you wish to believe that she or he were part of some reason to put all the heroes under your control, they do not have an ulterior motive."

This text is a clear example of [2], as the repetition of “suppose she never got the chance” doesn’t make any sense in the story. It also has a subtle grammatical error [1] of a comma being generated immediately after the period.

Example 5

Admittedly, the Ponte Building here is no Empire State Building. To be kind, it looks like a cement knockoff of the Capitol Records Tower in Los Angeles. But that's not what makes this building so special. The design of the building was by architect John D. Rockefeller Jr., who also designed the White House and the U.S. Supreme Court. The building was completed in 1892, just two years after the Great Chicago Fire.

This generation contains many factual errors ([5]). The Ponte Building is located in South Africa and was built in 1975. Even if you don't know these two facts, it should be obvious that the architect who designed the White House would be dead by 1892.

Getting Started

1. Please go to roft.io, choose login from the menu in the upper right, and then **sign up for an account**. Otherwise we can't keep track of how many times you've played the game and we can't give you credit for what you did!
2. When you begin, please play 20 rounds in a row in the same category. Here's which category to pick
    1. If your ID ends in 0 or 1, first play 20 rounds of the Short Story category
    2. If your ID ends in 2 or 3, first play 20 rounds of the Presidential Speeches category
    3. If your ID ends in 4 or 5, first play 20 rounds of the New York Times category
    4. If your ID ends in 6 or 7, first play 20 rounds of the Recipes category
    5. If your ID ends in 8 or 9, first play 20 rounds of the Random category
3. After you've played your first 20 rounds, you can pick any category that you'd like to try and continue playing.
4. You can check how many rounds of the game you have played by clicking on the profile link, and looking at the "Number of Games Played" stat.
5. You can earn extra points on the leaderboard by playing more rounds. So you're more likely to get some Extra Credit if you keep playing!