Do Massively Pretrained Language Models Make Better Storytellers?

Abigail See   Aneesh Pappu*   Rohun Saxena*   Akhila Yerukola*   Christopher D. Manning

*equal contribution

Our research questions

How does **massive pretraining** affect story generation?

- Large-scale pretrained Language Models have amazing performance on Natural Language **Understanding** tasks.
- But are they better at Natural Language **Generation** (NLG)?
- GPT2 has generated some amazing examples 🦄
  …but does it generate better text **in general**? Better **in what ways**?

How does the **decoding algorithm** affect story generation?

- Choice of **decoding algorithm** can greatly impact generated text.
- But many NLG papers only evaluate a **single** decoding algorithm (e.g. top-*k* sampling with one *k*). This gives an **incomplete view**.
- How does the generated text vary across **all** values of *k*?

Method

WritingPrompts dataset: a story generation dataset of stories (mean ~700 words) based on prompts (mean ~28 words).

We compare **two** language models:

- **The Fusion Model** (Fan et al, 2018): a convolutional seq2seq model designed for and trained on WritingPrompts.
- **GPT2-117** (Radford et al, 2019): a Transformer language model pretrained on WebText (unlabeled text corpus, ~45x size of WritingPrompts), then finetuned on WritingPrompts.

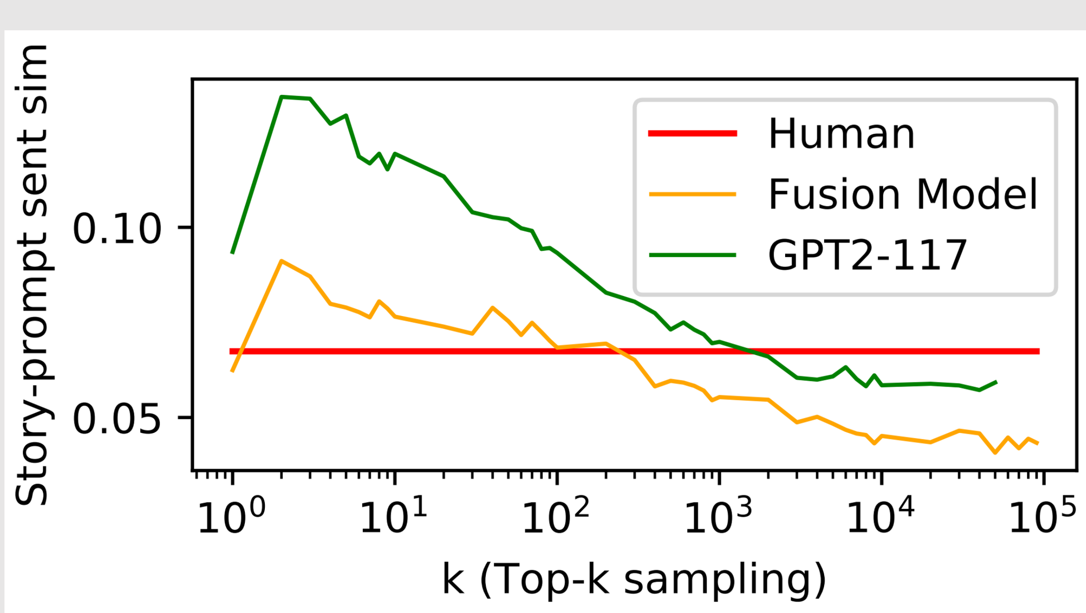
  Limitation: GPT2-117 is the smaller version of the full GPT2 model.

We use **top-*k* sampling** to generate stories **across the range** of *k*.

We use several **automatic** metrics to evaluate the stories.
Limitation: We do not use human evaluation.

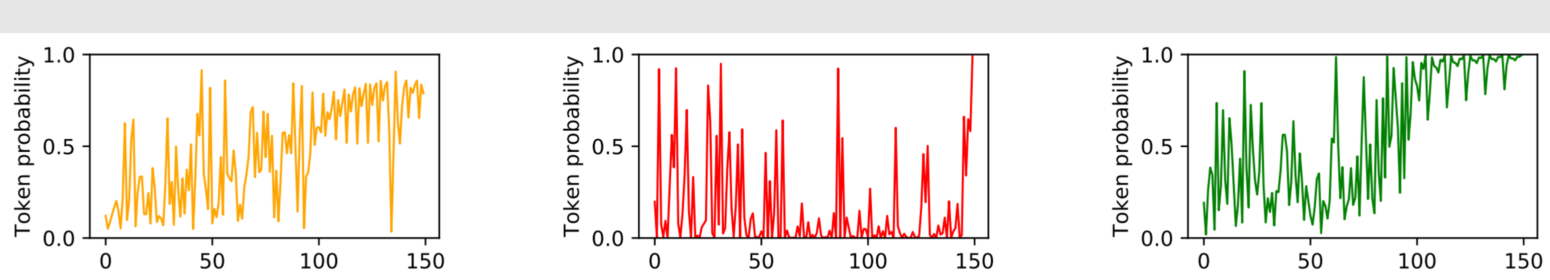
Story-prompt relatedness

GPT2-117 **conditions on the prompt more strongly** than the Fusion Model, generating stories that are **more similar to the prompt**.



Model confidence

When *k* is small, both models (**left** and **right**) get stuck in **repetitive loops** with **increasing confidence** – unlike human text (**center**).



(a) **Fusion Model** ($k = 2$): *I had never seen a man so young before. I had never seen him before, but he had always seemed to be a man of a man. He was young, and he was young. He was a man of a man, and a man who was young, and a man who was* [...]

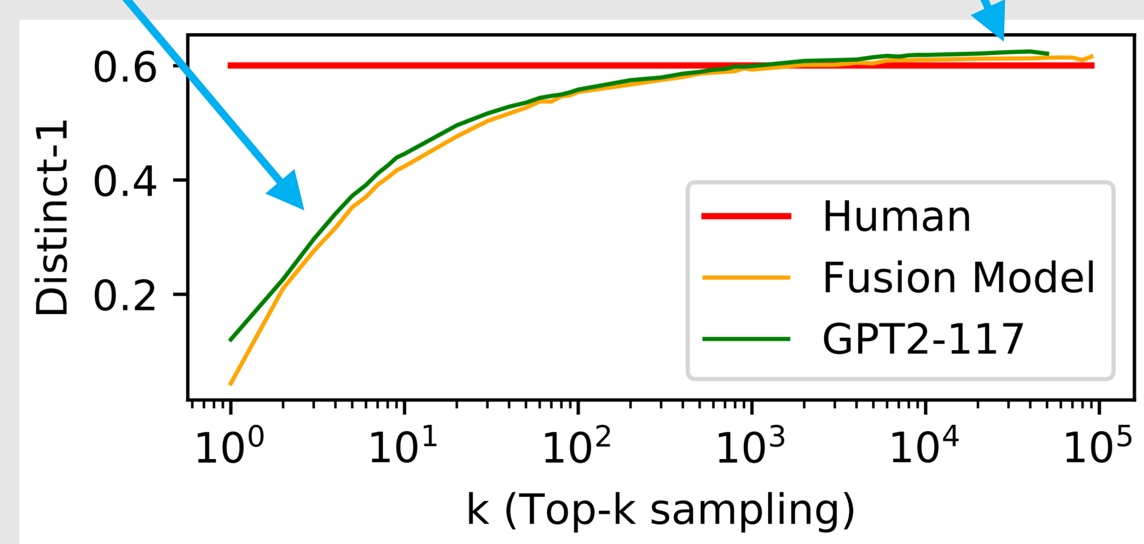
(b) **Human Text**: *"Looks like the rain's stopped." I peered out the window. Art was right; time to get to work. "Alright, let's move out." I could hear the scraping of the stone armor as the men slowly stood. Despite the training,* [...]

(c) **GPT2-117** ($k = 2$): *I've always been a man of the people. I've always been a strong man. I've always been a strong man. I was born in the city, I was raised in the country. I was raised in a family that wasn't very good. I 'm not a good man.* [...]

Repetition and Rareness

When *k* is small, both models generate **repetitive generic text**.

As k approaches vocabulary size, both models **converge to human levels** of repetition and rareness.



Repetition and genericness is mainly caused by choice of decoding algorithm, not a lack of training data!

Examples

When *k* is **small**, both models are **more repetitive and generic**

GPT2-117 stories are **more related to the prompt** than Fusion Model stories

When *k* is **small**, verbs are **less concrete** (e.g. *is*) and nouns are **more concrete** (e.g. *mother, father, queen*)

Prompt: You just discovered England's biggest secret: every time they say long live the queen it extends the queens life.

Human-Written Story: I am immortal, I have inside me blood of kings. My name in Charles Windsor and I am heir to a throne. Prince of Wales, they've called me that my entire life and now I lay on my death bed and my mother - still going strong. I'm not sure how it's possible but she seems ageless, almost timeless. My whole life I've waited for that [...]

Fusion Model ($k = 10$): *I have no idea when my father would come to visit this place, I was always so excited. I was always so scared. I had to go back to work. I was a little scared of him, but when I was younger, he had never come in for a visit. His mother told me, "I'm so scared", as he always did. She never talked to me about anything* [...]

GPT2-117 ($k = 10$): ** " So what are we going to do? " * It was the first time she asked me this question that I ever had asked her. It has taken years but finally I am finally able to do it. It seems that this queen is the queen of England and she's finally the one to tell me that this queen is the reason why she's not being allowed to die.* [...]

Fusion Model ($k = 1000$): *"Where am I? What happened?" "Having been working on my job for over 6 hours now, I do not know how you worked!" "I have been working for the last three years. Surely I am an ancient god now." The bar patrons snickered. "Hello?" "Those last three years have been worse than a year ago." Pain.* [...]

GPT2-117 ($k = 1000$): *It was an odd occasion for the Queen of England to meet with her. The plane sat idle at 3:26 PM on a Thursday night. Yesterday, the Queen had taken it upon herself to try and get a good look at the plane which had recently been found abandoned. A copious amount of curious glances from around the room until* [...]

When *k* is **large**, both models are **less repetitive and generic**

GPT2-117 uses **more named entities** (e.g. *England, Thursday*) than the Fusion Model

When *k* is **large**, verbs are **more concrete** (e.g. *meet, sat*) and nouns are **less concrete** (e.g. *pain, glances*)

Conclusions

The effect of **massive pretraining**?

- **The good:** GPT2-117 **conditions more strongly** on context, is more sensitive to **event ordering**, and generates text with **more concrete words** and **named entities** (compared to the Fusion Model).
- **The bad:** GPT2-117 is equally **repetitive, generic, syntactically under-complex**, and **over-confident** when *k* is small (compared to Fusion).
  These problems won't be solved by more training data!

The effect of ***k* in top-*k* sampling**?

When *k* is small, the models generate text that:

- is **repetitive, generic**, and uses a **smaller range of syntactic patterns**
- uses **more verbs and pronouns**, but **fewer nouns and adjectives**
- has **more concrete nouns** but **fewer concrete verbs**

These are side-effects of likelihood-maximizing decoding, not a fault in the models themselves!

also includes greedy and beam

When *k* is large, the models generate text that:

- **fits the patterns of human text** for most automatic metrics we measured
- …but is **nonsensical** and **lacks multi-sentence coherence**.

This work was funded by the Gerald J. Lieberman Fellowship, Tencent, and the DARPA CwC program under ARO prime contract no. W911NF15-1-0462.