

# MOTIVATION

In 2020 there was a virus that some considered to be the apocalypse .



" What do we do? " asked Irudim as he gestured to the two million persons in attendance, tired and shaken.

" Yeah, we die, " interjected Yuzsagm, " We get sick. "

Irudim opened his mouth, but his words did not catch, nor did the rest of the tal...



- Narrative generation (or story generation) is the task of generating a creative response given an input prompt
- Neural models and methods proposed to date for story generation and dialogue generation have been similar
- **How do existing decoding methods from neural response generation perform in narrative generation?**
- This work performs an empirical analysis on the effect of **nucleus sampling thresholds** and **maximum mutual information diverse decoding objective** (MMI-antiLM) on **narrative quality** (e.g. interestingness, coherence)

# FINDINGS

- Nucleus sampling is best between 0.7 and 0.9
- MMI objective improves the quality of generated stories ( $\lambda \leq 0.1$ )
- Established automatic metrics (distinct-n and sentence-BERT) do not correlate well with human judgments of narrative quality on any qualitative metric



READ THE PAPER!



JOIN THE BOT NARRATIVE GENERATION COMMUNITY  
[R/STORYGENAI](#)

# DECODING METHODS FOR NEURAL NARRATIVE GENERATION



Aaron Mueller\*  
Johns Hopkins University

Alexandra DeLucia\*  
Johns Hopkins University

João Sedoc  
New York University

Xiang Lisa Li  
Stanford University

## METHODS

### FINE-TUNE GPT-2 FOR NARRATIVE GENERATION

- Fine-tune small and medium **GPT-2** on creative writing dataset
- **r/WritingPrompts** from Reddit
- Fine-tune models on short, medium, and long versions of the narratives

### DECODER PARAMETER SWEEP

- Generate responses with GPT-2 conditioned on the prompts with various nucleus sampling and MMI parameters
- Nucleus sampling  $p$ : {0.0, 0.3, 0.5, 0.7, 0.9, 0.95, 1.0}
- MMI-antiLM  $\lambda$ : {0.0, 0.1, 0.2, 0.35, 0.5}

### EVALUATE WITH HUMAN AND AUTOMATIC METRICS

- Human evaluation on Amazon Mechanical Turk to judge **fluency**, **interestingness**, **coherence**, and prompt **relevance** (4-point Likert scale)
- Automatic metrics of **sentBERT** and **distinct-n** to evaluate diversity

## EXAMPLE GENERATION

The following was generated with GPT-2 Medium fine-tuned on the medium-length WritingPrompts dataset. Decoding parameters: nucleus sampling 0.7 with various anti-LM MMI decoding thresholds.

**Prompt: Your baby starts crying every time you leave their room . Finally , you pick up your child and leave the room together . As soon as you step out, you hear crying coming from the room .**

“Daddy!”

$\lambda=0.00$

I cried into my mom's breast. She was holding my son in her arms. I 'd never seen a baby cry so deeply. Tears were streaming down my face. I ran to the door and opened it to see a crying girl. I ran back to the room and opened the door, and there was my little girl. She 'd started crying again.

“What is wrong, mom?”

$\lambda=0.10$

There were only two of us left in the hospital. The nurses wouldn't let me leave. I was in pain. My mother was looking at me with a confused look on her face.

“Can you tell me what's wrong, honey?” she asked.

“Honey, it's okay.” I whispered. I couldn't hear her, and the nurses couldn't hear me either.

$\lambda=0.50$

“Daddy? Daddy what's wrong honey? Daddy why are you crying honey? Daddy pick up your child and leave the room, you have to get to work”

I picked up my daughter and we walked out into the kitchen.

I held her close and whispered into her ear “It's ok honey, I'll be ok.”