frontiers
in Computer Science

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fcomp.2021.674333

# Design and Analysis of a Collaborative Story Generation Game for Social Robots

*Eric Nichols[1]*, Leo Gao[2], Yurii Vasylkiv[3] and Randy Gomez[1]*

*[1]Honda Research Institute Japan Co., Ltd., Wakoshi, Japan, [2]EleutherAI, Edmonton, AB, Canada, [3]HCI Lab, University of Manitoba, Winnipeg, MB, Canada*

Storytelling plays a central role in human socializing and entertainment, and research on conducting storytelling with robots is gaining interest. However, much of this research assumes that story content is curated. In this paper, we introduce the task of *collaborative story generation*, where an artificial intelligence agent, or a robot, and a person collaborate to create a unique story by taking turns adding to it. We present a collaborative story generation system which works with a human storyteller to create a story by generating new utterances based on the story so far. Our collaborative story generation system consists of a publicly-available large scale language model that was tuned on a dataset of writing prompts and short stories, and a ranker that samples from the language model and chooses the best possible output. We improve storytelling quality by optimizing the ranker's sample size to strike a balance between quality and computational cost. Since latency can be detrimental to human-robot interaction, we examine the performance-latency trade-offs of our approach and find the optimal ranker sample size that strikes the best balance between quality and computational cost. We evaluate our system by having human participants play the collaborative story generation game and comparing the stories they create with our system to a naive baseline. Next, we conduct a detailed elicitation survey that sheds light on issues to consider when adapting our collaborative story generation system to a social robot. Finally, in a first step towards allowing human players to control the genre or mood of stories generated, we present preliminary work on steering story generation sentiment polarity with a sentiment analysis model. We find that our proposed method achieves a good balance of steering capability and text coherence. Our evaluation shows that participants have a positive view of collaborative story generation with a social robot and consider rich, emotive capabilities to be key to an enjoyable experience.

**Keywords: storytelling, interactivity, language models, AI agents, social robotics**

OPEN ACCESS

***Edited by:***
*Giovanni Luca Christian Masala,*
*Manchester Metropolitan University,*
*United Kingdom*

***Reviewed by:***
*Yan Wu,*
*Institute for Infocomm Research*
*(A ∗ STAR), Singapore*
*Umberto Maniscalco,*
*Institute for High Performance*
*Computing and Networking (ICAR),*
*Italy*

****Correspondence:***
*Eric Nichols*
*e.nichols@jp.honda-ri.com*

***Specialty section:***
*This article was submitted to*
*Computer Vision,*
*a section of the journal*
*Frontiers in Computer Science*

***Received:*** *01 March 2021*
***Accepted:*** *22 July 2021*
***Published:*** *25 October 2021*

***Citation:***
*Nichols E, Gao L, Vasylkiv Y and Gomez R (2021) Design and Analysis of a Collaborative Story Generation Game for Social Robots.*
*Front. Comput. Sci. 3:674333.*
*doi: 10.3389/fcomp.2021.674333*

## 1 INTRODUCTION

Storytelling is a central part of human socialization and entertainment. Many of the popular forms of storytelling throughout history–such as novels, plays, television, and movies–have passive audience experiences. However, as social robots become more widespread, they present a new avenue for storytelling with a higher level of interactivity. Much research has been dedicated to issues surrounding with how robots should connect with an audience (Mutlu et al., 2006; Gelin et al.,

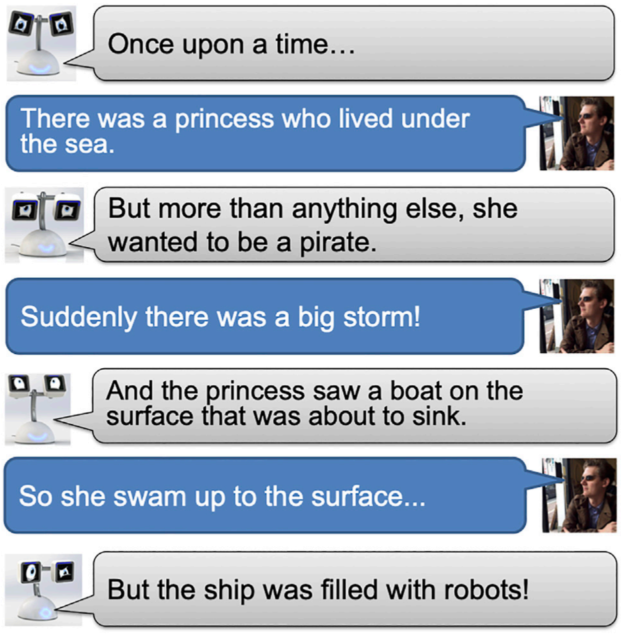

Once upon a time…
There was a princess who lived under the sea.
But more than anything else, she wanted to be a pirate.
Suddenly there was a big storm!
And the princess saw a boat on the surface that was about to sink.
So she swam up to the surface...
But the ship was filled with robots!

**FIGURE 1 |** Collaborative storytelling with an intelligent agent.

**TABLE 1 |** Example story starters from `r/WritingPrompts`.

| Story starters |
|---|
| *When the human race arrived on the galactic scene, nobody was even remotely impressed* |
| *All eyes rested on the tall man dressed in black as he sat silently, staring into nothingness with a intensity that seemed out of place in the massive cathedral* |
| *James walked down the row of rooms; they did not call them prison cells, since the occupants were not prisoners* |
| *I took a deep breath, closed my eyes, and waited for the world to return to life* |
| *The Big Bad Wolf worked for the bank, and he had to evict the pigs who were delinquent on their mortgages* |
| *In one hand John held what would most certainly be his death, and in the other, he held a message* |
| *King Klotopan let a raised fist fall, and with a roar of the crowd the match began* |
| *The Historical Revision Agency was created on June 16th, 2048, following the invention of working time travel 6 mo previously* |
| *The giggling children, pushing and laughing with one another, began excitedly pointing up toward the softly whistling orb* |
| *None of us had really expected senior skip day to play out like this* |

2010; Ham et al., 2015; Costa et al., 2018; Gomez et al., 2018; Ligthart et al., 2020), but so far, most work works under the assumption that story content would be curated in advance. At the same time, recent breakthroughs in language modeling present a new opportunity: language, and thus stories, can potentially be generated on demand.

In this paper, we introduce a novel game of *collaborative story generation*, where a human player and an artificial intelligence agent or robot construct a story together. The game starts with the AI agent reciting one of a curated set of *story starters*–opening sentences meant to kick-start participants' storytelling creativity–and the human player responds by adding a line, which we refer to from here on out as a *story continuation*, to the story. The AI agent and human player then take turns adding to the story until the human player concludes it. The game is designed to have a few restrictions as possible and contrasts with traditional storytelling settings where the narrative is fixed in advance.

Collaborative story generation builds on a rich tradition of collaboration in storytelling that includes Dungeons and Dragons, improvisational comedy, and theater. It also bears some resemblance to *Choose Your Own Adventure* style novels[1] where users make narrative choices to progress the story. collaborative story generation could be a useful tool for encouraging creativity and overcoming writer's block, as well as being an entertaining game in its own right.

Our end goal is to make it possible for intelligent agents, such as robot companions and avatars (Park et al., 2019; Gomez, 2020; Gomez et al., 2020), to play the collaborative story generation game, as shown in **Figure 1**. The examples throughout this paper come from real stories that were constructed by humans collaborating with a text interface version of our storytelling system[2].

Our contributions are as follows:

1. We introduce the task of *collaborative story generation*, where a human and an intelligent agent or social robot construct a story together and consider the arising technical and presentational challenges.
2. We present a *sample-and-rank*-based approach to *collaborative story generation* that combines a large-scale neural language model with a sampler and ranker to maximize story generation quality.
3. We conduct rigorous analysis of our approach to fully understand its performance and the trade-offs between quality and latency, and we use these findings to determine the optimal number of story continuation candidates to generate.
4. We conduct qualitative evaluation including evaluation of stories generated by our system in isolation, and two distinct populations of human judges of actual stories constructed by humans playing with our collaborative story generation system. The evaluation confirmed our optimized ranker model's contribution to story generation quality.
5. We conduct a detailed elicitation survey to gain insight into potential user preferences for a collaborative story generation game with the social robot, Haru (Gomez, 2020). Survey findings were positive overall and highlighted the importance of the robot conveying the emotional contents of the stories.
6. Finally, with the goal of allowing players to specify the mood or genre of the story during collaborative story generation, we present preliminary work on steering story generation using a sentiment analysis model showing we could successfully influence story generation sentiment without degrading quality.

---

[1] https://en.wikipedia.org/wiki/Choose_Your_Own_Adventure

[2] Stories were edited for brevity.

**TABLE 2** | Example stories from `r/WritingPrompts`. Story starters are shown in red.

| **Story 1** |
|---|
| *Conventions are my favorite time of the year.* We all knew we were killers, and knowing who all the other killers were helped people get along. Sold a secret? Everyone would hunt you down the next day. Everyone had their flair out, so you could have as much fun as you wanted while still fitting in. The Japanese would dress all in earth tones, while the Italians would put on fantastically gauche suits. The British, as always, sent their best in a tuxedo. I wore my best denim. Nobody ever expected a Canadian assassin. We were there in WWII, killing Hitler in his bunker before the other Allies even arrived. They had to burn the body and claim suicide to cover up their incompetence. We were there when Osama bin Laden was found. I'll bet the Navy Seals won't tell you they found him chained to his desk with a complimentary last meal of poutine. Don't think we're friendly. We don't take anybody's side but Canada's. We were there when JFK was shot |
| **Story 2** |
| *I wouldn't call it a war.* Extermination maybe. Though I'd more aptly describe it as a harvest. By the time they reached our world and penetrated the stratosphere, people sought them out in droves to be harvested. Of course, they knew what that actually meant. Otherwise, they wouldn't have been so eager. Ten years before the Angels descended from the sky, they had already sent what some referred to as divine retribution: a virus. Though this virus in particular only targeted women. It spread faster than a wildfire and had a 100% mortality rate. Worse yet, it was completely undetectable. In our desperation, we became animals. We locked our wives, daughters, and mothers deep underground under the constant shine of UV radiation and still they got infected. Within 5 yr, the last woman had died leaving the rest of humanity to slowly die with her |
| **Story 3** |
| *Andrew nearly snarled as his phone chimed for what seemed to be the umpteenth time.* Andrew rubbed a hand over eyes to clear them, trying to understand the ridiculous amount of messages but he must have been more tired than he thought since they didn't make sense. He quickly scrolled through the message previews, finding they all seemed to follow the same pattern: to look at the moon. The moon? What the -? Why? It was like a shot to his adrenaline when he saw that some of the messages were coming from Isabel's phone. He shot up from bed, seeing the other side empty, and jumped to his feet. "Hun, what's going on," Andrew questioned, still scrolling through his phone while walking towards the guest room. The room was at the end of the hallway and the door was wide open. He could see Isabel standing in the middle of the room, arms down by her side and phone clutched in her hand as she gazed out the window. As he grew closer, he could *see* that she was shaking, "Bel? Honey? What's wrong, wh-" Isabel's body whipped toward him and Andrew couldn't help himself, he froze in place. She ran and shoved her body into the door, slamming it shut and locking it in place |

## 2 COLLABORATIVE STORY GENERATION

In this section, we describe our formulation of the collaborative story generation game. The goal of this game is to provide human players opportunity to get creative and express themselves. We do so by providing an intelligent agent that plays along with humans, loosely following the *yes-and* principle that was loosely codified in improvisational theater (Johnstone, 2012): the intelligent agent's contribution should follow the premises set up by the story up to the current point.

The collaborative story generation game flows as follows:

1. The intelligent agent selects and recites a story starter from our curated collection: these are catchy opening lines meant to kickstart the storytelling process with an interesting premise. Examples are given in **Tables 1**, **2**.
2. The human player responds by adding a line to the story. There are no restrictions, and the human player can have the story continue however they want.
3. The intelligent agent follows by reading the story up to the current point and adding a line that seems likely to follow the human contribution.
4. The collaborative story generation game continues by alternating between steps *2*) and *3*) until it concludes.

There are many possible strategies that could be used to conclude the collaborative story generation game: the human player could simply declare *The End*; the intelligent agent could decide based on player engagement levels, the flow of the story, or other information that is it time to generate an ending; or after deciding to end the story, the agent could cheer on the human player on to develop an ending.

Story ending generation is an important research topic that has grown in interest recently (Zhao et al., 2018; Guan et al., 2019; Luo et al., 2019), and it is beyond the scope of this paper. We plan to address this in future work but limit the scope of the collaborative story generation presented here to a fixed number of exchanges between human player and intelligent agent in order to simplify evaluation.

We design the collaborative story generation game to have as few restrictions on the human player as possible. In principle, our storytelling system can handle input of arbitrary length and content, and stories can last as long as the human player wants. In practice, the mode of interaction does impose some constraints: when playing the game with a text-based interface, story generation quality will degrade if spelling and punctuation are not in agreement with the system's conventions. When playing the game verbally, the quality of the robot's speech recognition can have a similar impact. We plan to thoroughly explore these issues in future work, but at this stage we do not constrain the players' input.

## 3 MATERIALS AND METHODS

Our approach to collaborative story generation is as follows: a *generator model* generates story continuations given the text of the story so far, a *sampler* uses the generator to produce a specified number of story continuation candidates, and a *ranking model* or *ranker* picks the best possible choice as the system's contribution. Combining a language model with a ranker allows adjusting the trade-off in computational costs and generated story quality. Our collaborative story generation architecture is shown in **Figure 2**.

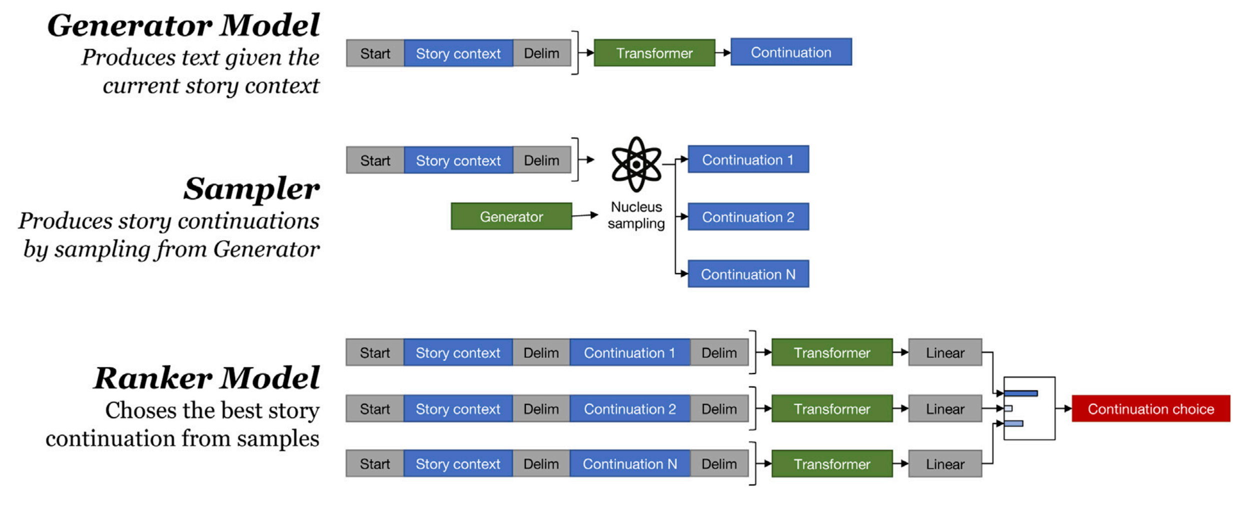


**FIGURE 2 |** Our collaborative storytelling system architecture.

**TABLE 3 |** An example of the collaborative story generation process.

| | **Story context:** |
|---|---|
| Prompt: | *I shot the guy behind me a smile, wanting to believe he was laughing at my inaudible joke* |
| Human: | *But I knew he wasn't just the kind of guy to laugh at people* |
| System: | *But then he whispered something in my ear and the smile left me* |
| Human: | *"Meet me out back behind the alley, and bring the money." he said* |
| | `<|endoftext|>` |
| **Ranker Score** | **Candidate story continuations:** |
| 0.169 | *"Put it down I want my money."* |
| 0.006 | *I gave him the money as a matter of fact, sat in the alley and made my way to the alley he left behind* |
| 0.006 | *The guards had a stone shard in one of their hands and had raked the money out* |
| 0.115 | *When I came back to the alley a few hours later, I found a motorcycle about a mile away* |
| 0.069 | *"I'm late for work" I said and punched in a fake name before quickly exiting* |
| 0.200 | *I felt a rush of adrenaline and turned around* |
| 0.204 | *I wish I knew what was going on* |
| 0.146 | *The lights in the alley flickered but they always went out when someone came close* |
| 0.037 | *That was how I got out of that date with Jimmy Jim* |
| 0.048 | *I started walking around looking in and out of the dark and suddenly I noticed it and saw what looked like a car stop* |
| System: | *I wish I knew what was going on* |

*When generating the next continuation in a story, the story generation system receives as input the story context – a concatenation of all lines in the story, including the story starter, human's utterances, and system-generated continuations. The story generation model then uses Nucleus sampling to generate a set number of candidates from the Generation model and ranks them with the Ranker model. The storytelling model then outputs the highest-scoring story continuation, and the game continues.*

To play the collaborative story generation game using the system, the game is started by randomly selecting a story started from the WritingPrompts data set described in **Section 3.4.2**. Then at each step where a story continuation is required, the language model reads the entire story up to that point including utterances from both the human player and system and the specified number of story continuation candidates is generated. Then the sampler and ranker are used to choose the best possible continuation. This process continues until the game concludes. An example of the story generation process is given in **Table 3**.

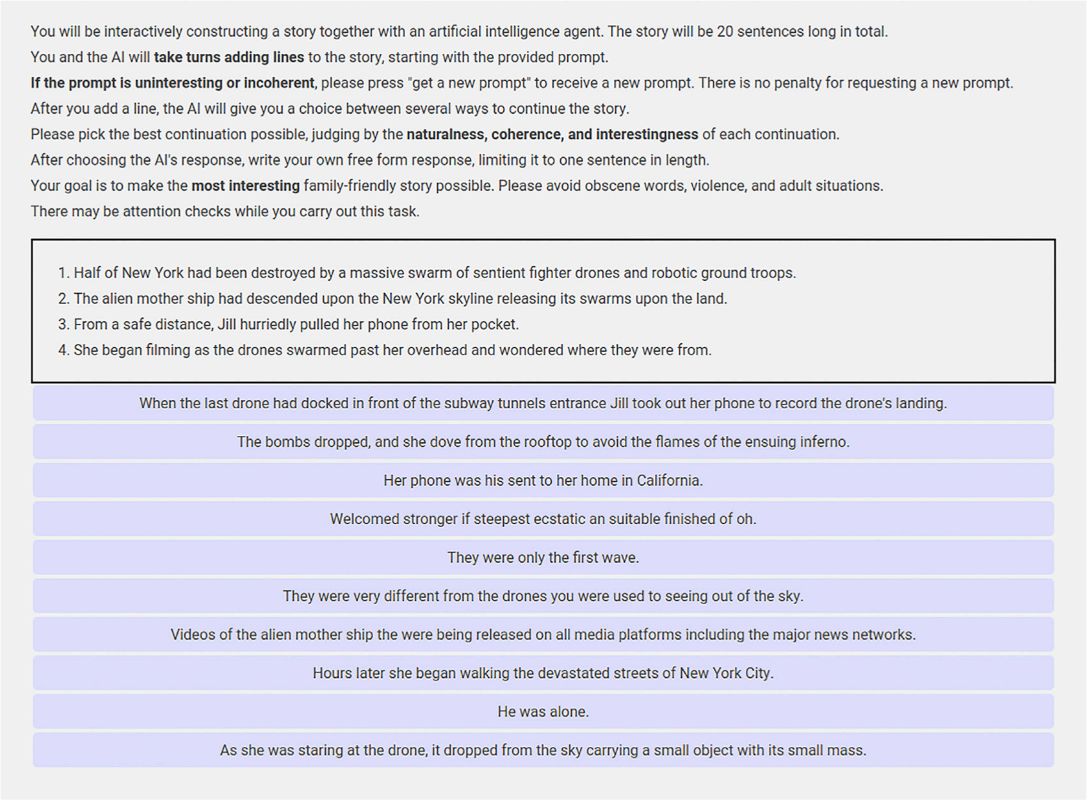
## 3.1 Generator Model

The Generator is a unidirectional autoregressive language model which is sampled from multiple times to generate candidate story continuations. We used the publicly-available pretrained 774M parameter GPT-2-large model[3] (Radford et al., 2019).

One issue with using a language model for generation is the output may be ill-formed or lacking in logical coherence. The main solutions for this issue are the use of larger models, the use of different sampling methods, and the use of various methods of traversing the search space of possible sentences. However, larger models are at greater risk of over-fitting and result in large increases in memory usage for modest gains in quality, which makes them impractical to use. As such, we focused on sampling and searching through ranking.

---

[3] https://github.com/openai/gpt-2



You will be interactively constructing a story together with an artificial intelligence agent. The story will be 20 sentences long in total.
You and the AI will **take turns adding lines** to the story, starting with the provided prompt.
**If the prompt is uninteresting or incoherent**, please press "get a new prompt" to receive a new prompt. There is no penalty for requesting a new prompt.
After you add a line, the AI will give you a choice between several ways to continue the story.
Please pick the best continuation possible, judging by the **naturalness, coherence, and interestingness** of each continuation.
After choosing the AI's response, write your own free form response, limiting it to one sentence in length.
Your goal is to make the **most interesting** family-friendly story possible. Please avoid obscene words, violence, and adult situations.
There may be attention checks while you carry out this task.

1. Half of New York had been destroyed by a massive swarm of sentient fighter drones and robotic ground troops.
2. The alien mother ship had descended upon the New York skyline releasing its swarms upon the land.
3. From a safe distance, Jill hurriedly pulled her phone from her pocket.
4. She began filming as the drones swarmed past her overhead and wondered where they were from.

When the last drone had docked in front of the subway tunnels entrance Jill took out her phone to record the drone's landing.
The bombs dropped, and she dove from the rooftop to avoid the flames of the ensuing inferno.
Her phone was his sent to her home in California.
Welcomed stronger if steepest ecstatic an suitable finished of oh.
They were only the first wave.
They were very different from the drones you were used to seeing out of the sky.
Videos of the alien mother ship the were being released on all media platforms including the major news networks.
Hours later she began walking the devastated streets of New York City.
He was alone.
As she was staring at the drone, it dropped from the sky carrying a small object with its small mass.

**FIGURE 3 |** Web interface for collaborative storytelling annotation task. Participants select from amongst ten possible story continuations generated by the system before adding their own line to the story. Reproduced with permission from (Nichols et al., 2020).

## 3.2 Sampler

The most popular approaches for sampling from autoregressive models have predominantly focused on techniques for truncating the low-quality tail of the model distribution, like top-k and nucleus sampling (Holtzman et al., 2019). Sampling is used in most GPT-2 based text generation systems, superseding greedy or untruncated sampling. In all experiments, we use nucleus sampling with $p = 0.9$.

## 3.3 Ranking Model

The Ranker model scores each story continuation candidate and selects the highest scoring one. It is a standard GPT-2-large model with a final classification head consisting of a linear layer outputting a single scalar for each token. The input format to the model is: `(context)<|endoftext|>(choice)<|endoftext|>`.

The `<|endoftext|>` token is used because it is guaranteed not to occur elsewhere in the input. As GPT-2 is unidirectional, the embedding of the final token integrates information from the entire input context window; this is similar to the use of the `[CLS]` token in BERT. Thus we execute the Ranker model once for each choice, keep only the outputs from the last token of the final layer for each choice as the logit score of each choice, and compute a softmax over them. The Ranking model architecture is shown in **Figure 2**.

We chose a neural network-based ranker model to select the best story completion from the Generator output because it offers us control over the trade-off between text generation quality and computational demand, while avoiding the significantly increased memory footprint and inflexibility in computational cost of using a larger language model. The amount of computational resources used is easily adjustable by changing the number of rollouts considered by the Ranker. This serves as a middle ground between the intractable extreme of searching the entire space of all vocab$^{\text{length}}$ possible sentences, and the computation-efficient but suboptimal solution of sampling without any branching or backtracking.

One popular alternative search solution making a similar trade-off is beam search, which keeps a dynamic list of generation candidates. Beam search has been applied in many language generation tasks, including machine translation (Tillmann and Ney, 2003). However, sampling from a language model using beam search can lead to degenerate text (which is typically repetitive and uninteresting), in an open-ended task such as storytelling (Holtzman et al., 2019). These issues are avoided using a neural network-based ranker model because it has richer text representations, it scores full text utterances rather than incomplete text fragments, and it can incorporate additional information about the storytelling domain from its training data.

## 3.4 Datasets

In this section we describe our datasets: *1*) a collaborative story generation dataset constructed by crowdsourcing workers

**TABLE 4 |** Example stories generated by the `tuned` system with a human through the collaborative storytelling annotation task.

| **Human Chat** `tuned` **Story 1** | |
|---|---|
| Prompt: | *A man sits alone at a table, listening to the low and pleasant hum of a bassist pluck something smooth and jazzy nearby* |
| Human: | *He lets out a big sigh and dreamily looks at the bassist, sipping on his whisky* |
| System: | *The music is not too loud, just enough to drown out his worries, but still enough to cause him to feel like he has heard it before* |
| Human: | *The man tries his best to just immerse himself into the music until it ended* |
| System: | *But then the phone rang, making the man jump out of his chair* |
| Human: | *"Darn it," says the man as he frantically searches his pockets for the phone* |
| System: | *As the man was getting ready to answer, a beautiful lady, with long flowing blonde hair, greeted him* |
| Human: | *The man drowned out the constant ringing on the phone and silenced it, saying hello back* |
| System: | *She turns to the guy and smiles while staring him straight in the eyes* |
| Human: | *The man replied "My name is Jacob…how about you"?* |
| System: | *After a few moments of silence, she smiled and explained that the man had already lived with her for years* |
| Human: | *"Huh…what do you mean I already live with you?" Jacob said in utter confusion* |
| System: | *She explained that she was the same age as he, and is the love of his life* |
| Human: | *"No, you got to be kidding me I'm single and I come to this bar every day," Jacob said with a quiver in his voice* |
| System: | *"You know what's funny?"* |
| Human: | *"Please explain, and are you just a stranger playing a joke on me?"* |
| System: | *She handed him a picture of her and a girl* |
| Human: | *Her thumb was covering another person, and she slowly uncovered it as Jacob looked on* |
| System: | *"Look at me," said the girl* |
| Human: | *"This…picture shows me, you, and a child…but why?" Jacob started to tear up* |
| System: | *"Just a moment, it will all make sense soon enough."* |

| **Human Chat** `tuned` **Story 2** | |
|---|---|
| Prompt: | *The garage door lurched open and my face turned from joy to agonizing surprise* |
| Human: | *A small squirrel leaped from behind my car, likely as shocked to see me as I was him* |
| System: | *"I'm so sorry, I didn't mean to scare you."* |
| Human: | *He tilted his head to the side, almost indicating he understood my words* |
| System: | *"You don't scare me."* |
| Human: | *Completely floored that this squirrel can speak English, I fumbled myself into asking "Are you real? This can't be real!"* |
| System: | *"Yes, this is real, and I have been waiting for you for quite some time."* |
| Human: | *I asked him "What have you been waiting for? Some peanuts from the kitchen?!"* |
| System: | *"No, I can answer all your questions with just one sentence!"* |
| Human: | *He leaned against the tire of my car with a sly look on his face, "I'm madly in love with Penny, your cat, and I'd like to take her to dinner tonight!"* |
| System: | *I sat there confused for a good 30 s before I snapped out of it* |
| Human: | *"Have you been spying on my house cat!?"* |
| System: | *The squirrel cocked his head to the side and shrugged, as if he had no response to what I said* |
| Human: | *"Well I'll have you know that I am not okay with you romantically pursuing my poor Penny!"* |
| System: | *This time, the squirrel cracked his knuckles and the look in his eyes made me question everything I was hearing* |
| Human: | *"Look Mr, all I want to do here is show her a good time and maybe be a part of the family!"* |
| System: | *I was dumbfounded* |
| Human: | *I felt a slight lump in my throat realizing that maybe this squirrel was lonely and just wanted a good home and asked "Well, I guess if you insist, where did you want to take her to dinner?"* |
| System: | *"To my favorite fishing spot!"* |
| Human: | *The squirrel chuckled, pulled out a teeny tiny fishing rod, slapped on a little fishing hat and started making his way inside to ask my soon to be very confused cat out on a date* |
| System: | *Thank you for reading!* |

*Reproduced with permission from (Nichols et al., 2020).*

interacting with our collaborative story generation system that are used to train the Ranker model and for evaluation, and *2*) a writing prompts dataset comprised of short stories written in response to writing prompts posted to a Web forum that are used to train the Generator model.

### 3.4.1 Collaborative Story Generation Dataset

We collected collaborative stories using Mechanical Turk, each consisting of 20 interactions in response to a provided story starter (which is sampled from the initial sentences of stories in the WritingPrompts dataset described in **Section 3.4.2**). The interactions in the story alternate between *choice* type interactions, in which a human participant chooses from 10 story continuations that are generated by out collaborative story generation system, and *freeform* type interactions, in which the human participant is able to provide a complete sentence response. The Web interface for this task in shown in **Figure 3**.

In order to ensure data quality, one of the continuations in the *choice* type interaction is a *distractor* which is made by concatenating randomly sampled words. The distractors are also filtered through Mechanical Turk beforehand by asking workers whether the sentences are coherent or not, and only the ones labelled incoherent by workers are used. As a quality check, if a worker selects a distractor during a *choice* type interaction, the story is discarded.

We collected a total of 2,200 stories, which we randomly partitioned into a training split of 2,000 stories, and validation and test splits of 100 stories each. Some example stories generated by human participants together with our system are shown in **Table 4**.

### 3.4.2 Writing Prompts Dataset

We constructed a dataset of stories from the `r/WritingPrompts` subreddit[4], consisting of all posts with score greater than 3 made before 2019-11-24, amounting to 140 k stories in total. Some heuristics were used to clean the stories[5]. This data was used to train the Generator model. Example stories are given in **Table 2**.

We chose to collect our own WritingPrompts dataset instead of using the FAIR WritingPrompts dataset (Fan et al., 2018), because it gave us the flexibility to filter stories by custom score thresholds, as well as to perform the different preprocessing necessary for GPT-2. Our dataset also contains more than an additional year's worth of data compared to the FAIR dataset.

## 3.5 Story Continuation Sampling and Ranking

To generate story continuations from our system, sentences are generated from the Generator model and filtered using a set of cleanliness heuristics until the desired number of samples is achieved. Our heuristic rejected sentences with less than 60% alphabetic characters, unbalanced quotations, select profanity, or words like "chapter" that are not typically part of the story. The Ranker model then computes a score for each story continuation and selects the highest scoring one.

## 3.6 Training

### 3.6.1 Generator Model

The Generator model is trained with a maximum likelihood estimation loss function using Adafactor (Shazeer and Stern, 2018) with a learning rate of $5e-5$ on a weighted mixture of the WritingPrompts and BookCorpus (Zhu et al., 2015) datasets. The addition of BookCorpus helps reduce the risk of over-fitting on the comparatively smaller WritingPrompts dataset.

### 3.6.2 Ranking Model

The Ranking model is trained on the WritingPrompts dataset and eight copies of the training split of the collaborative story generation dataset, shuffled at the story level. Each batch for the Ranking model consists of 20 sentences taken from a single story. To ensure that the model fits in memory, only the sentences that fit within 400 tokens are used, resulting in some batches with less than 20 sentences. The majority of stories do not have to be truncated.

The Ranker model is also trained on synthetic collaborative story generation data that we create from the WritingPrompts dataset. Stories with less than 100 characters or 35 sentences are first removed from the Ranking model's training data. Then the first sentence of the story is used as the story starter, and the next 20 sentences are all used as the preferred story continuations of *choice* type interactions, where the other nine incorrect choices are sampled from the 25th and subsequent sentences of the story.

The Ranking model is trained using Adam (Kingma and Ba, 2014) with a maximum learning rate of $1e-5$. The entire model is trained; no layers are frozen. The checkpoint is resumed from a GPT-2 text generation model that was tuned on the BookCorpus and WritingPrompts datasets in the same way as the Generator model.

## 3.7 Sentiment Steering

Preliminary human evaluation of our collaborative story generation system highlighted participant's desires to influence the mood and genre of stories, but such explicit control of language generation has been a challenge for language models. However, in recent years there have been developments in *language model steering*, where judgements from a classifier are directly incorporated into a LM's training and/or generation. For example, generated text can be steered toward a target domain (e.g. Wikipedia- or tweet-like text) (Dathathri et al., 2020) or target emotion; obscenities or other undesirable text can be filtered out (Keskar et al., 2019).

Krause et al. (2020) proposed GeDi, a method for language model steering that combines the probability distribution over categories from a classification model with the probability distribution over the next token to be generated by an LM. This has the advantage that no retraining of the LM is necessary, and any classifier can be applied. However, GeDi and similar approaches often suffer from degraded text generation quality, if the classification model exerts too much influence over generation.

Inspired by GeDi (Krause et al., 2020), we propose a sentence-level approach to language-model steering that can benefit from an external classification model while avoiding text degradation. We do so by combining our Ranker model with a classification model to select sentences that are more likely to share target categories. Because our goal is to allow human collaborative story generation players to specify story moods or genres, as an initial trial we use a sentiment classification model to steer text generation towards either `positive` or `negative` sentiment.

Thus, we need a sentiment analyzer that can perform well on text generated during collaborative story generation. To keep the architecture of our system simple and preserve efficiency by requiring only GPT-2 based LMs, we train our own sentiment model. Our sentiment model is a 124M parameter GPT-2-small model with a classification head, similar to the Ranking model, and trained to producing a score, where 1.0 is interpreted as `positive` and 0.0 in interpreted as `negative`. We trained our model on the TweetEval (Rosenthal et al., 2017; Barbieri et al.,

---

[4] https://www.reddit.com/r/WritingPrompts/

[5] We removed smart quotes, links and user/subreddit mentions, and all HTML entities and markdown formatting.

**TABLE 5 |** Examples from the TweetEval (Rosenthal et al., 2017; Barbieri et al., 2020) dataset.

| Sentiment | Tweet |
|---|---|
| Positive | To prepare for the #AppleEvent tomorrow I am reminiscing over legacy apple hardware including iPod 2nd Gen, iPhone 4, iPad 2. Get Pumped! |
| Positive | 5 Days off now! Bit of golf on the agenda and a few other bits. Looking forward to watching @user Ant-Man on Thursday too! |
| Positive | Today is National Ice Cream Day!!! Give in to the excuse! LOL. |
| Negative | Yes that's right @user close down your best channel why don't you? You're not meant to be Netflix or Amazon Prime. |
| Negative | Briana just sat her ass there while Matt said he's done with her sister? This chic! She deserves everything she gets! #LittleWomenLA |
| Negative | "if the #iPhone5 doesn't come out on September 7th, I'm suing Steve Jobs...I can't put up with the Motorola Razr much longer" |

*Positive and Negative labels are color-coded accordingly.*

**TABLE 6 |** Comparison of the sentiment analysis accuracy of several state-of-the-art transformer-based sentiment analysis systems.

| System | Architecture | Dataset | Accuracy (%) |
|---|---|---|---|
| A | BERT | Yelp reviews | 58 |
| B | DistilBERT | IMDB movie reviews | 68 |
| C | BERT | IMDB movie reviews | 72 |
| D | BERT | Stanford Sentiment Treebank 2 | 76 |
| E | BERT | Stanford Sentiment Treebank 2 | 77 |
| F | RoBERTa | TweetEval | 78 |
| G | GPT-2 | TweetEval | 78 |

*Models were downloaded from Huggingface. Our approach, System G (`rank_sent`), achieves comparable performance to the best-performing publicly-available models.*

2020) dataset, a collection of Twitter annotated at the tweet level with `positive` and `negative` labels. Examples from the TweetEval dataset are show in in **Table 5**. Training was carried out using Adam (Kingma and Ba, 2014) with a maximum learning rate of $1e-5$.

To verify that our sentiment analyzer has sufficient performance on collaborative story generation text, we compare it to existing sentiment analyzers on a small evaluation dataset we construct by having Mechanical Turk workers annotate 100 randomly-selected sentences from our collaborative story generationdataset described in **Section 3.4.1** with sentiment polarity judgements. Three workers annotate each sentence, and we use majority voting to select the correct label. We then evaluated the performance of a small number of transformer-based sentiment analyzers that were trained on different datasets.

The results are shown in **Table 6**, where our approach is labeled system G (`rank_sent`). We can see that models trained on Twitter data (systems F, G) outperform models trained on restaurant reviews (system A) or movie reviews (systems B, C, D, E) and that architecture variants do not have much impact on performance. Our model performs comparably to the best-performing publicly-released model: a RoBERTa model that was trained on Twitter data, providing evidence that using our own GPT-2-based sentiment analyzer is unlikely to degrade performance.

To combine the sentiment analyzer score with the ranking model score, we use use Bayes theorem with the ranking model to provide a prior and updating with a weighted version of the sentiment model (with weight $\omega$, indicating our degree of confidence in the sentiment model). This scoring is similar to Krause et al. (2020) but acts at the sentence rather than token level. This computation is as follows:

$$f_{\text{score}}(x|\text{positive}) \propto f_{\text{ranker}}(x) f_{\text{sentiment}}(\text{positive}|x)^{\omega} \qquad (1)$$

There are several advantages of optimizing at sentence level instead of token level. As all sentiment models are only proxies for sentiment, by maximizing for the sentiment score, quality invariably suffers due to regressional Goodharting (Manheim and Garrabrant, 2018), a phenomenon where a proxy measure is optimized for instead of the true goal. By constraining the sentiment model to only affect generation at the sentence level and by choosing from a small number of possible continuations generated from the Generator rather than having control at every single token, the amount of optimization pressure exerted is significantly limited. Limiting the amount of optimization pressure exerted has previously been considered as a method of limiting regressional Goodharting in (Taylor, 2016). As a result, sentence level sentiment models are robust to a much wider range of sentiment weight values. As will be seen in the evaluation in **Section 4.6**, even when $\omega = \infty$, the generated text is still generally coherent.

## 3.8 Related Research

In this section, we summarize relevant research in story generation, collaborative language generation, language modeling, and storytelling with robots.

### 3.8.1 Story Generation

In recent years, the task of automatic story generation has gained a lot of attention. Fan et al. (2018) construct a corpus of stories and propose a hierarchical story generation model. Yao et al. (2019) approach the task by first generating a plot outline and then filling in the language. Gupta et al. (2019) generate story endings by incorporating keywords and context into a sequence-to-sequence model. Luo et al. (2019) incorporate sentiment analysis into story ending generation. See et al. (2019) conduct an in-depth analysis of the storytelling capabilities of large-scale neural language models. However, the primary assumption of these works is that story generation is conducted without human interaction.

### 3.8.2 Collaborative Language Generation

While research on collaborative language generation is still sparse, there are a few notable recent developments.

AI Dungeon[6] is a text adventure game that is generated by a GPT-2 language model (Radford et al., 2019) tuned on a collection of text adventure play-throughs. In the game, players assume the first person and interact with the world by inputting commands or actions. The language model is used to generate the world's reaction to the player's actions. Our *collaborative storytelling* task and approach are similar to AI Dungeon, but our task is not constrained to the genre of first-person adventures, and we perform ranking of model outputs.

Cho and May (2020) build an improvisational theater chatbot by identifying and collecting instances of improvisational dialogue on the Web and using it to tune and evaluate public domain dialogue systems. Our collaborative storytelling task is similar to improv, but stories are linguistically different enough from improv that it would be impractical to apply their dataset to our task. In addition, our approach employs sampling and ranking to improve the likelihood that language model utterances are in the desired storytelling domain, while Cho and May (2020) use the model's output as-is.

### 3.8.3 Language Models

In order for an AI agent to participate in collaborative storytelling, it must be able to generate story continuations. A language model (LM) is a mathematical model that assigns likelihoods to sequences of words where sequences that are more likely in a target language are given higher scores. Such a model can be used to generate text.

Early language models estimated token sequence likelihood based on token sequence counts taken from large collections of text together with various smoothing methods to handle novel token sequences (Ney et al., 1994). Later, RNNs and other sequential neural networks models became popular due to their ability to apply distributed word representations (Bengio et al., 2003; Mikolov et al., 2011; Sutskever et al., 2011), but RNNs have issues with vanishing gradients and modelling long-term dependencies found in text.

The recent transformer architecture (Vaswani et al., 2017) uses attention layers to model long-term dependencies by greatly increasing the model's visible context. Transformers have been shown to perform well in a variety of tasks, including machine translation (Vaswani et al., 2017) and a variety of language understanding (Radford et al., 2019) and language generation tasks (Zhang et al., 2019). A notable transformer model is BERT (Devlin et al., 2018). However, as it is a bidirectional model, BERT and its variants are rarely used for text generation, due to the necessity for computationally-expensive Gibbs sampling (Wang and Cho, 2019).

The model we use as a basis for our system, GPT-2 (Radford et al., 2019), is a large-scale neural network using the transformer architecture. GPT-2 is a general purpose unidirectional LM trained on a large corpus which has been successfully applied to many downstream tasks.

**TABLE 7 |** Accuracy of the `tuned+ranked` model at predicting the story continuation that was selected by the Mechanical Turker who constructed the story.

| System | Dataset | Accuracy (%) | |
|---|---|---|---|
| `tuned+ranked` | validation | 22.9 | (229/1000) |
| `tuned+ranked` | test | 23.3 | (233/1000) |
| *random baseline* | - | 10.0 | - |

*Note that a random baseline would pick the correct continuation 1 out of 10 times. Reproduced with permission from (Nichols et al., 2020).*

Recently, there has been interest in steering language model generation toward desired domains or styles, to avoid undesirable language, or to influence the emotion or mood of the text. Conditional Transformer Language Model ((Keskar et al., 2019), CTRL) and Generative Discriminator Guided Sequence Generation ((Krause et al., 2020), GeDi) incorporated classification models into a transformer-based LM to allow steering toward a target class at the token level. Plug and Play Language Model ((Dathathri et al., 2020), PPLM) introduce a series of control codes that can be inserted into a transformer-based LM at the beginning of generation to steer generation toward a target class. Token-level LM steering has the advantage that the target class can influence every token in the generation, however, this can often lead to degradation in text quality. In contrast, we develop a sentence-level LM steering method that uses ranking to allow the target class to influence generation while preserving text generation quality.

### 3.8.4 Storytelling With Robots

Storytelling with robots is a topic that has been heavily researched. *See* (Chen et al., 2011) for a survey. Much of recent research focuses on issues of connecting with audiences through gaze (Mutlu et al., 2006), gesture (Gelin et al., 2010; Ham et al., 2015), and emotive storytelling (Costa et al., 2018; Gomez et al., 2018), or a combination thereof (Wong et al., 2016; Wu et al., 2017). Other research investigates strategies of directly engaging with listeners (Ligthart et al., 2020).

Finally, (Sun et al., 2017) investigated joint storytelling with children participants, where a child participant created a story by selecting from a predefined set of characters and settings, a human experimenter made suggestions on story continuations following a limited number of engagement strategies, and a puppeteered robot later helped child participants recall the stories they created in order to evaluate the efficacy of the engagement strategies. Our task setting is different in that it focuses on having an AI agent automatically generating story continuations without restrictions in settings or content rather than on evaluating strategies for humans to engage with other human storytellers.

## 4 RESULTS

In this section, we present evaluation of our collaborative storytelling system. First we describe our two survey

---

[6] https://aidungeon.cc

**TABLE 8** | Mean acceptability of story continuations in the test set.

| System | Acceptability (%) | |
|---|---|---|
| `untuned` | 33.9 | (305/900) |
| `tuned` | 39.8 | (358/900) |
| `tuned+ranker` | 62 | (62/100) |

*To evaluate* `untuned` *and* `tuned`*, acceptability is calculated over all 9 continuations from each system, while* `tuned+ranked` *uses the Ranker to consider only the best one. Reproduced with permission from (Nichols et al., 2020).*

populations: *1*) a group of crowdsourcing workers, and *2*) a group of university students. Next, in order to verify our ranker's performance we evaluate it though story continuation acceptability and human annotator story preferences with our group of crowdsourcing workers. Then, in order to understand the trade-offs between story generation quality and latency, we investigate our ranker model. After, in order to gain insights into the characteristics that people feel our system has, we adapt the ACUTE-EVAL chatbot evaluation metric (Li et al., 2019) to collaborative storytelling evaluation and compare our proposed approach to a baseline by analyzing the story preferences of our survey groups. Finally, in order to gain insights into adapting interactive storytelling to a social robot, we conduct an elicitation survey with the second survey group.

## 4.1 Participants

The participants in this study consisted of two groups. The first group was recruited on Amazon Mechanical Turk (MTurk), and the second group were students from a local university in Winnipeg, Canada. One hundred participants were recruited from MTurk ($N$ = 100) and twenty two participants from the university ($N$ = 22). In the group of the university students were 11 males and 11 females with age ranging from 20 to 40 yr old (M = 28.5, SD = 2.1). Participants' backgrounds were not collected.

## 4.2 Story Continuation Acceptability

Story continuation prediction acceptability measures the accuracy of the Ranker model at predicting the continuation chosen by the Mechanical Turk worker that interacted with the model to produce the story. This metric is a proxy for how often the `tuned+ranked` picks the best continuation of the story, but its usefulness is diminished by variance in human annotators and the possibility of multiple equally good continuations. The results are summarized in **Table 7**. Nonetheless, we find that our Ranker model outperforms chance by a factor of over two, providing evidence that it is able to capture the preferences of human annotators to an extent.

As an additional measure of our systems' capacity to generate story continuations that match human preferences, we formulate the story continuation acceptability task. In this task, each story continuation generated by a system is classified as either *acceptable* or *unacceptable*, and we compare their mean acceptability precision.

We annotated the acceptability of candidate story continuations by asking Mechanical Turk workers to classify each continuation given the context of the story generated so far. To ensure annotation quality, we have three workers evaluate each *choice* interaction per story from both the validation and test sets and take the majority vote across the three labels as the final label[7]. These *choice* interactions consist of nine story continuations generated by the system and one incoherent distractor. If a worker labels a distractor acceptable, their annotations are discarded. We use this method to evaluate how often each model produces outputs that are an acceptable continuation of the story, rather than the best continuation.

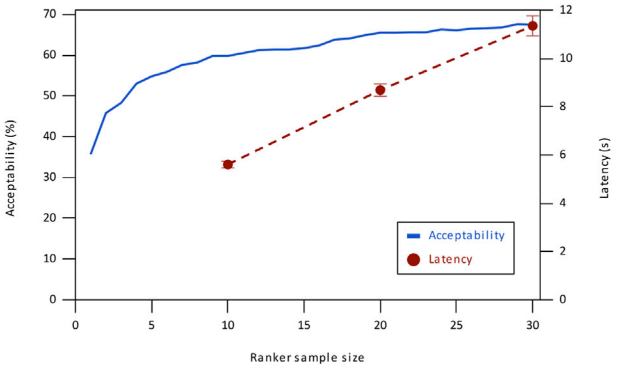


**FIGURE 4** | Ranker sample size vs. acceptability and latency. Error bars show standard error.

Since the `tuned` and `tuned+ranked` systems use the same language model samples, we use the test set to evaluate their performance, considering the mean acceptability of all of the sampled continuations from `tuned` and the acceptability of the single continuation selected by `tuned+ranked` for each *choice* interaction in the datasets. To evaluate the `untuned` system, we gather and evaluate 100 *choice* interactions by having Mechanical Turkers construct stories with the `untuned` system.

The results are summarized in **Table 8**. As we can see, the `tuned` system outperforms the `untuned` system, showing that tuning the language model on storytelling data is important in improving generation quality. We also find that `tuned+ranked` greatly outperforms the other two systems, providing supporting evidence that our Ranking model is effective at helping our language model produce story continuations that are likely to be preferred by humans.

## 4.3 Ranker Performance Trade-offs

In order to fully understand the trade-offs between quality and latency and ensure our collaborative storytelling model is fast enough to work with a robot agent, we investigate the

---

[7] The workers reached unanimous agreement 41.9% of the time on the test data.

**TABLE 9 |** Questions asked to human evaluators of collaborative storytelling systems.

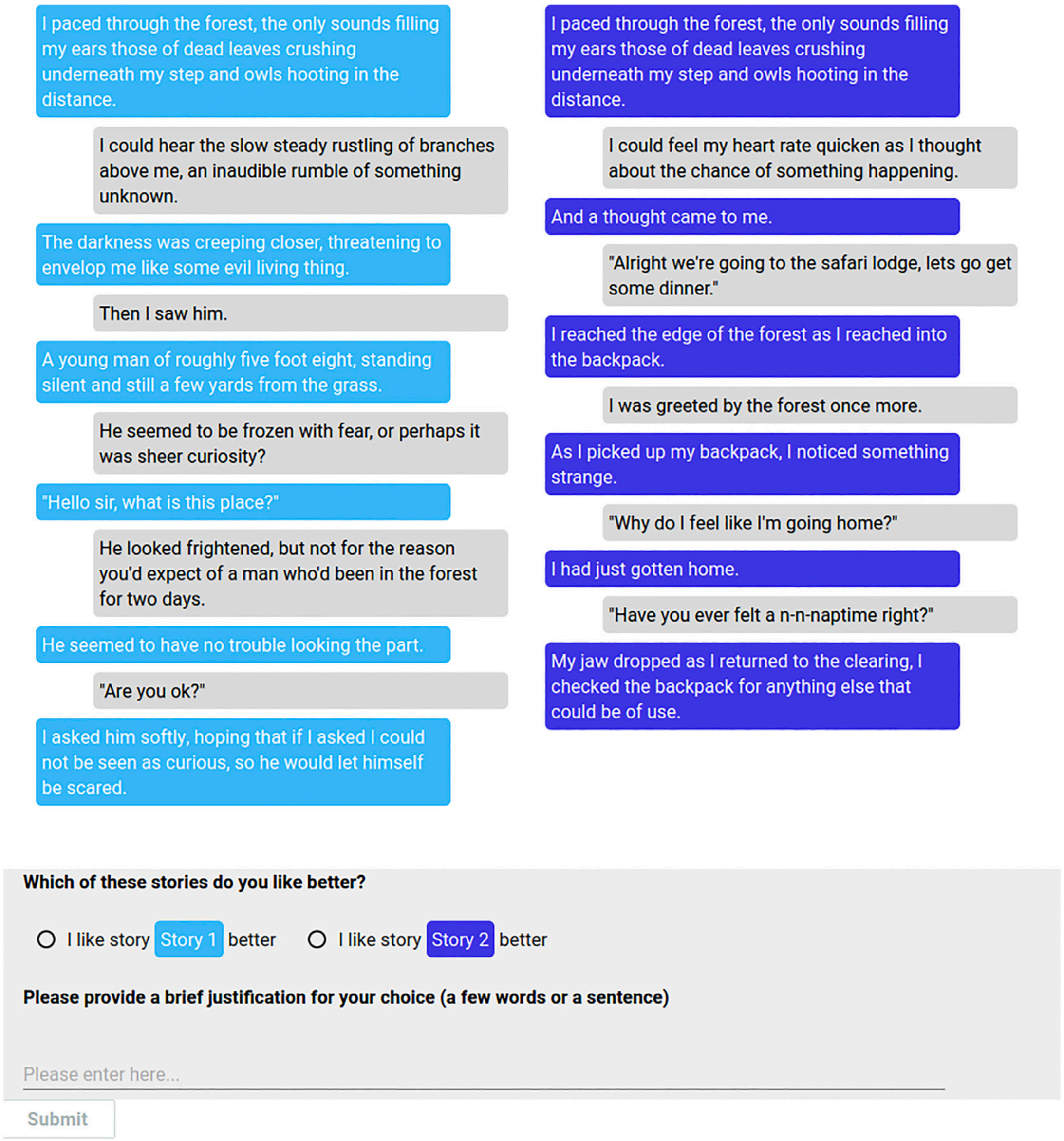
| Characteristic | Question |
|---|---|
| Engagingness | Who would you prefer to collaborate with for a long story? |
| Humanness | Which storyteller sounds more human? |
| Interestingness | If you had to say one of these storytellers is interesting and one is boring, who would you say is more interesting? |
| Story Preference | Which of these stories do you like better? |
| Repetitiveness | If you had to say one of these storytellers is repetitive and the other is not, who would you say is more repetitive? |
| Target Domain | Which of these seems more like a story? |
| Target Sentiment | If you had to say one of these storytellers is `<target-sentiment>` and the other is `<other-sentiment>`, who would you say is more `<target-sentiment>`? |

*Characteristics and questions are based on our modified-for-storytelling* ACUTE-EVAL *metric. Reproduced with permission from (Nichols et al., 2020).* Repetitiveness, Target Domain, *and* Target Sentiment *are additional characteristics used to evaluate sentiment steering-based approaches.*

**TABLE 10 |** Example stories generated by self-chat with the `tuned+ranked` system.

| **Self-chat `tuned+ranked` story 1** | |
|---|---|
| B: | *I stood before the immense bronze doors, arms crossed, awaiting the fate of the two evil Dictators* |
| A: | *"Ladies and gentlemen of the land" a soft yet authoritative voice boomed through the doors* |
| B: | *"We are here for your arrival" a commanding and officious voice echoed through the halls* |
| A: | *A slow breath tickled my lips as I was finally allowed inside* |
| B: | *The archway was brimming with people, many in red, with metal cufflinks, and most in ornate black garb* |
| A: | *I stepped forward into the main audience hall, pausing to take in my surroundings* |
| B: | *The ceiling was made of rough jagged glass panels; out in the back stood eight shiny and elaborately painted metal ladders* |
| A: | *The stately mahogany wooden podium was suspended from the ceiling, very tall and proud* |
| B: | *The four different Dictators circled the podium with their various groups in the room* |
| A: | *The oldest of the Dictators stood, hands together in front of him, on his magnificent golden podium* |
| B: | *He seemed more imposing than the rest, holding a bony silver sword in his shining gold hand* |
| A: | *"Hallowed be the name of his majesty, King Gautir" the King boomed through the open archway* |
| B: | *The King was a corpulent, middle aged man with his long white hair growing out of his crown, nearly the size of his small face* |
| A: | *I could see that his eyes were piercing like the flaming sun* |
| B: | *The King's long, white robes seemed long like a black robe worn by someone with a wide girth* |
| A: | *His sneering smile was reflected in his large, deep blue eyes* |
| B: | *"Today, we greet you, members of the committee, from the Holy Empire of All Flame and Sound."* |
| A: | *The King's voice boomed as the rest of the Dictators turned and stepped towards the podium, stopping on their footsteps* |
| B: | *"Welcome to this evening's meeting" the King announced to us* |
| A: | *"You two can either choose to remain in your chambers or come out and face the Committee" The King said in a commanding tone* |
| **Self-chat `tuned+ranked` Story 2** | |
| A: | *Have you ever wondered what exactly goes into making our favorite foods?* |
| B: | *You know, the big baked potato with oil and salt, the slushy that's savory enough to eat for lunch every day?* |
| A: | *The answer is simple: the food industry* |
| B: | *It's a fairly big industry, you see* |
| A: | *It's estimated that five hundred different corporations control 70% of the world's food supply* |
| B: | *If you were to think of it in simple terms, that's right: almost everyone on the planet eats food produced by a food corporation* |
| A: | *As you might imagine, the world is full of giant corporations and a huge amount of money being made through the food industry* |
| B: | *We've all heard the expression "money talks, money does."* |
| A: | *What do you think happens when that money goes through the food industry?* |
| B: | *It goes into the mouths of politicians and it goes into the mouths of stockholders* |
| A: | *This whole industry is very complicated, although people never seem to come up with any clear-cut answers for it* |
| B: | *But, really, the only thing that people seem to have a handle on about the food industry is that it's a big financial mess* |
| A: | *When you think about it, this seems pretty clear* |
| B: | *Let me explain* |
| A: | *In order to make money, companies have to produce enough food to feed a population* |
| B: | *When a company creates more food than it can eat, it has to buy more food, so that it can keep producing more food* |
| A: | *More food means more food prices, so that the corporation can make more money* |
| B: | *So how does the food industry achieve this?* |
| A: | *The answer to this is pretty simple* |
| B: | *In the world of food production, companies come up with marketing schemes that manipulate people's tastes* |
| A: | *Using a carrot for example, a company might work to improve a persons' reaction to carrots* |

*Reproduced with permission from (Nichols et al., 2020).*



I paced through the forest, the only sounds filling my ears those of dead leaves crushing underneath my step and owls hooting in the distance.

I could hear the slow steady rustling of branches above me, an inaudible rumble of something unknown.

The darkness was creeping closer, threatening to envelop me like some evil living thing.

Then I saw him.

A young man of roughly five foot eight, standing silent and still a few yards from the grass.

He seemed to be frozen with fear, or perhaps it was sheer curiosity?

"Hello sir, what is this place?"

He looked frightened, but not for the reason you'd expect of a man who'd been in the forest for two days.

He seemed to have no trouble looking the part.

"Are you ok?"

I asked him softly, hoping that if I asked I could not be seen as curious, so he would let himself be scared.

I paced through the forest, the only sounds filling my ears those of dead leaves crushing underneath my step and owls hooting in the distance.

I could feel my heart rate quicken as I thought about the chance of something happening.

And a thought came to me.

"Alright we're going to the safari lodge, lets go get some dinner."

I reached the edge of the forest as I reached into the backpack.

I was greeted by the forest once more.

As I picked up my backpack, I noticed something strange.

"Why do I feel like I'm going home?"

I had just gotten home.

"Have you ever felt a n-n-naptime right?"

My jaw dropped as I returned to the clearing, I checked the backpack for anything else that could be of use.

**Which of these stories do you like better?**

○ I like story Story 1 better ○ I like story Story 2 better

**Please provide a brief justification for your choice (a few words or a sentence)**

Please enter here...

**FIGURE 5 |** The Web interface for comparing stories for self-chat and human chat evaluations with ACUTE-EVAL. Reproduced with permission from (Nichols et al., 2020).

optimal number of story continuation candidates to generate.

### 4.3.1 Ranker Sample Size

We analyze the effect of varying the number of choices presented to the Ranker on the mean story continuation acceptability metric presented in the previous section.

Results are shown in **Figure 4**. While increasing the number of choices considered by the Ranker has the greatest effect on Acceptability between 0 and 10 choices, we observe a slower but continued improvement in quality throughout. The continued monotonic improvement indicates that the Ranker model is robust and prefers better continuations even when given a larger number of choices.

### 4.3.2 Latency

We measure the latency of our system for various numbers of Ranker choices to ensure that latency is acceptable. Our results are shown in **Figure 4**. We measure the latency using a single 1080Ti for the Generator model and with the Ranker model run on CPU due to insufficient GPU memory. We take the mean across 100 different stories. For each story, we only look at the latency for the final completion to provide an approximate upper bound, as latency increases towards the end of a story.

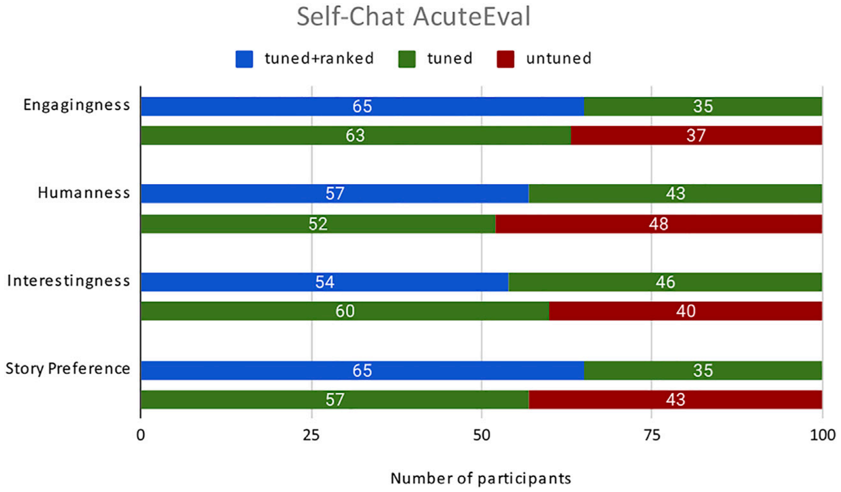


**FIGURE 6 |** Human evaluation of self-chat stories, comparing (`tuned+ranking`, `tuned`) and (`tuned`, `untuned`) pairs of systems. The preferred system is indicated by a larger portion of the bar.

### 4.4 Story Preference Evaluation

Conducting qualitative evaluation of collaborative storytelling is challenging for several reasons. Due to the subjective nature of stories, it is challenging to automatically evaluate their quality, and many methods have to make due with indirect measures such as entropy that offer limited insights, or ending generation where example desired text is known. In addition, the highly interactive nature of the task means that the influence of human participants makes it difficult to isolate the performance of the system. Here we present two forms of evaluation. To gain an understanding of our storytelling system's performance in isolation, we conduct *self-chat* evaluation of stories produced by conversing with itself. To understand participants' subjective perfection of the storytelling task, we evaluate *human chat* stories constructed by humans engaging in the collaborative storytelling task with our system.

#### 4.4.1 Evaluation Metric

Instead, since collaborative storytelling involves language exchange between entities with turn taking, we take inspiration from dialogue system evaluation methodology. Faced with the challenge of comparing multiple dialogue systems, Li et al. (2019) developed a method of comparing conversation pairs that instructs evaluators to only pay attention to the contributions of a single specified speaker in the conversation. In addition, their evaluation method, known as ACUTE-EVAL, allowed them to evaluate the contributions of a given dialogue system in terms of characteristics, such as engagingness, interestingness, humanness, and knowledgeability. Finally, to evaluate different dialogue systems without requiring a human to chat with them, they apply the self-chat technique of Ghandeharioun et al. (2019) and generate conversations for evaluation by having dialogue systems talk to themselves.

We create our own evaluation metric based on the characteristics targeted by the *PersonaChat* metric of ACUTE-EVAL[8]. In this evaluation, participants are asked to compare two stories produced through collaborative storytelling with different systems and to express their preferences in a series of questions. For each target characteristic, we take the question that Li et al. (2019) identified as most likely to differentiate between the evaluation of two systems and reword it to fit the collaborative storytelling setting. Finally, we add a question to measure overall story preference. The questions we ask, designed to help us gain insight as to which systems are easier and more enjoyable to make stories with, are shown in **Table 9**.

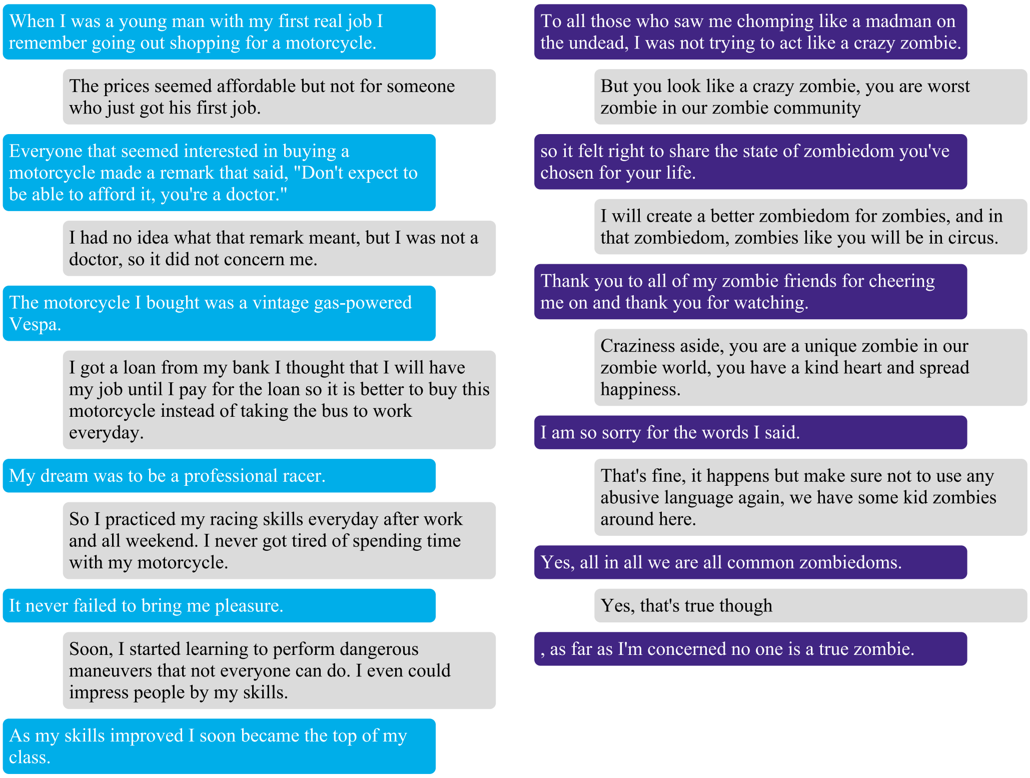
#### 4.4.2 Self-chat Evaluation

To eliminate variance from human storytellers, we use the self-chat setting of Li et al. (2019), where each model converses with itself. Some example stories are shown in **Table 10**. We compare the `untuned` and `tuned+ranked` models against the `tuned` model. For each pair of models, we collect 100 comparisons per question, and we instruct workers to provide short justifications for their decisions. The Web interface shown to workers is given in **Figure 5**.

The results of the evaluation are summarized in **Figure 6**. For each characteristic evaluated, the pairs of models are shown as stacked bar graphs, where a larger portion represents a stronger preference for that system. As can be seen, `tuned` is preferred over `untuned`, and `tuned+ranked` is preferred over `tuned` for all characteristics and overall story preferences, providing evidence that tuning the language model on storytelling

---

[8] We exclude the *Wizard of Wikipedia* metric because knowledgeability is not directly relevant to our collaborative storytelling setting.



When I was a young man with my first real job I remember going out shopping for a motorcycle.

The prices seemed affordable but not for someone who just got his first job.

Everyone that seemed interested in buying a motorcycle made a remark that said, "Don't expect to be able to afford it, you're a doctor."

I had no idea what that remark meant, but I was not a doctor, so it did not concern me.

The motorcycle I bought was a vintage gas-powered Vespa.

I got a loan from my bank I thought that I will have my job until I pay for the loan so it is better to buy this motorcycle instead of taking the bus to work everyday.

My dream was to be a professional racer.

So I practiced my racing skills everyday after work and all weekend. I never got tired of spending time with my motorcycle.

It never failed to bring me pleasure.

Soon, I started learning to perform dangerous maneuvers that not everyone can do. I even could impress people by my skills.

As my skills improved I soon became the top of my class.

To all those who saw me chomping like a madman on the undead, I was not trying to act like a crazy zombie.

But you look like a crazy zombie, you are worst zombie in our zombie community

so it felt right to share the state of zombiedom you've chosen for your life.

I will create a better zombiedom for zombies, and in that zombiedom, zombies like you will be in circus.

Thank you to all of my zombie friends for cheering me on and thank you for watching.

Craziness aside, you are a unique zombie in our zombie world, you have a kind heart and spread happiness.

I am so sorry for the words I said.

That's fine, it happens but make sure not to use any abusive language again, we have some kid zombies around here.

Yes, all in all we are all common zombiedoms.

Yes, that's true though

, as far as I'm concerned no one is a true zombie.

**FIGURE 7 |** Examples stories by univ. students and collaborative storytelling systems: `ranker30` **(left)** and `untuned` **(right)**.

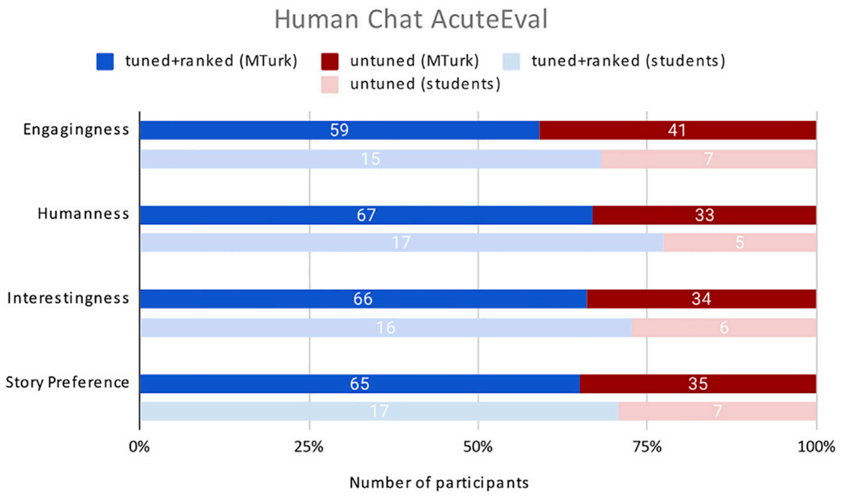


**FIGURE 8 |** Human evaluation of stories created with human collaborators, comparing our selected ranker `ranker30` (here labeled `tuned+ranked`) to an `untuned` baseline. The preferred system is indicated by a larger portion of the bar.

data and ranking the generated story continuations make complementary contributions to our collaborative storytelling system's performance.

### 4.4.3 Human Chat Evaluation

In order to gain a better understanding of the influence that human participants can have on collaborative storytelling results,

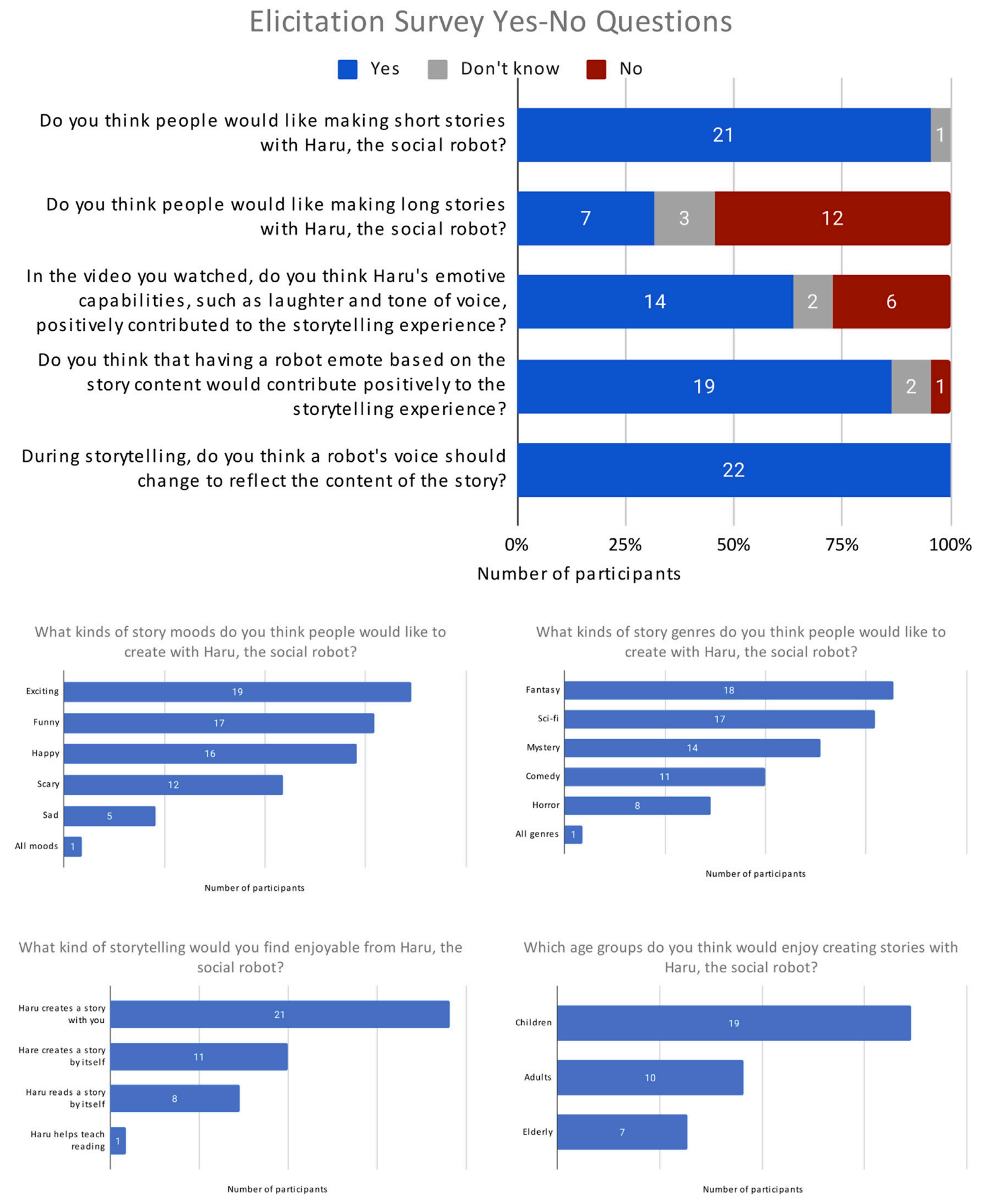


**FIGURE 9 |** Questions and responses from the elicitation survey.

we conduct and additional evaluation in the human chat setting, where the stories evaluated are generated by human participants from the two groups described in **Section 4.1** using a Web interface to our collaborative story generation generation system. The example interface for comparing stories is shown in **Figure 5**. Example stories created by participants and used in evaluation are shown in **Figure 7**. Results of ACUTE-EVAL are given in **Figure 8**.

Our ranking model is shown to be preferred over a baseline on every characteristic by both groups of participants,

**TABLE 11 |** Sentiment steering settings compared.

| Setting | System | Unit | Weight ($\omega$) |
|---|---|---|---|
| `rank_sent_low` | Sentence ranking | Sentence | 1 |
| `rank_sent_medium` | Sentence ranking | Sentence | 3 |
| `rank_sent_high` | Sentence ranking | Sentence | 9999 |
| `gedi_token_low` | GeDi | Token | 10 |
| `gedi_token_medium` | GeDi | Token | 30 |
| `gedi_token_high` | GeDi | Token | 1000 |
| `baseline` | N/A | N/A | N/A |

with statistically-significance differences in most cases[9], reinforcing findings from our self-chat evaluation that tuning and ranking a large-scale language model performs better at storytelling than an untuned model, and suggesting that our optimization of the Ranking model further improved performance.

## 4.5 Elicitation Survey

In this section, we present results from an elicitation survey designed to give insight into adapting our collaborative storytelling system for use in a robot.

As our target robot, we selected the social robot Haru (Gomez et al., 2018; Gomez, 2020), due to its rich, emotive capability. Haru is an experimental tabletop robot for multimodal communication that uses verbal and non-verbal channels for interactions. Haru's design is centered on its potential to communicate empathy through richness in expressivity (Gomez et al., 2018). Haru has five motion degrees of freedom (namely base rotation, neck leaning, eye stroke, eye rotation and eyes tilt), that allows it to perform expressive motions. Furthermore, each of the eyes includes a 3-inch TFT screen display in which the robot eyes are displayed. Inside the body there is an addressable LED matrix (the mouth). Haru can communicate *via* text-to-speech (TTS), through animated routines, projected screen, etc. Haru's range of communicative strategies positions the robot as a potent embodied communication agent that has the ability to support long-term interaction with people.

We conduct the elicitation survey with the university student group of participants described in **Section 4.1**. The crowdsourcing workers were excluded due to challenges associated with taking lengthy surveys on the MTurk platform.

Survey participants are shown a video simulation of a human playing the collaborative storytelling game with Haru. After completing the video, they answer a series of questions covering topics ranging from potential target demographics to emoting in robot storytelling. The questions and their responses are shown in **Figure 9**. We summarize our findings below.

On potential target demographics, `children` was overwhelmingly the most suggested age group (19 responses), with `adults` and `elderly` suggested less than half as often.

For potential storytelling modes, collaborative storytelling ("Haru creates a story with you") was the most suggested storytelling mode for Haru (22 responses); alternatives received a maximum of 11 responses. One participant suggested having Haru teach reading. We also found that participants overwhelmingly prefer making short stories to long stories (21 `yes` responses vs. 8 `yes` responses respectively), suggesting long storytelling sessions could be tiring for participants.

For storytelling content, positive moods were far more popular than negative: `{exiting,funny,happy}` received 52 cumulative votes vs. `{sad,scary}`, which received 17 cumulative votes. These results may be reflective of children's popularity as a suggested target demographic. Genre preferences showed a similar trend: `{fantasy,mystery,sci-fi}` (49 cumulative votes) were more popular than horror (eight cumulative votes), though the effect was not as strong with comedy (11 cumulative votes). These results suggest that human players may enjoy having control over mood or genre during collaborative story generation.

In a hypothetical scenario, participants greatly prefer an emotive robot (17 votes) to a disembodied voice (three votes), and want the robot's voice to reflect the content of the story (22 `yes` votes).

Finally, regarding emoting in storytelling, an overwhelming majority of participants think a robot emoting based on story content would contribute positively to the storytelling experience (19 `yes` votes), but there is room for Haru's emoting to improve, as few participants felt Haru's emotive performance contributed positively (14 `yes` votes).

These survey results suggest that collaborative storytelling with Haru is a potentially enjoyable application, but they also reinforce recent findings (Mutlu et al., 2006; Ham et al., 2015; Gomez et al., 2020) that emphasize the importance of convincing emotive delivery in robot storytelling. Thus, we must carefully consider Haru's emotive delivery during collaborative storytelling for it to be impactful.

## 4.6 Sentiment Steering

In this section, we investigate the effectiveness of sentence-level sentiment steering for collaborative story generation by comparing our proposed approach (`rank_sent`) to the token-level steering approach of GeDi (`gedi_token`) (Krause et al., 2020).

### 4.6.1 Evaluation Settings

For evaluation, we use Salesforce's publicly-released implementation of GeDi. To enable direct comparison, we use the same story generation model for each approach: the `tuned` GPT-2 language model from the self-chat evaluation in **Section 4.4.2**. However, since `gedi_token` requires a token-level sentiment analyzer, we use the GPT-2-based sentence-level sentiment analyzer we created in **Section 3.7** for `rank_sent` and GeDi's default token-level sentiment analyzer for `gedi_token`.

Since LM steering approaches combine the predictions of a classification model with the LM, choosing an appropriate weight for the classifier is essential to ensure a good balance. Too large of a

---

[9] We test for significance using a two-sided binomial test with null hypothesis $\mu_0 = 0.5$. The Interestingness, Humanness, and Story Preference metrics achieve significance at the $p < 0.005$ level on the MTurk evaluation. The Humanness and Story Preference metrics achieve significance at the $p < 0.05$ level on the university evaluation.

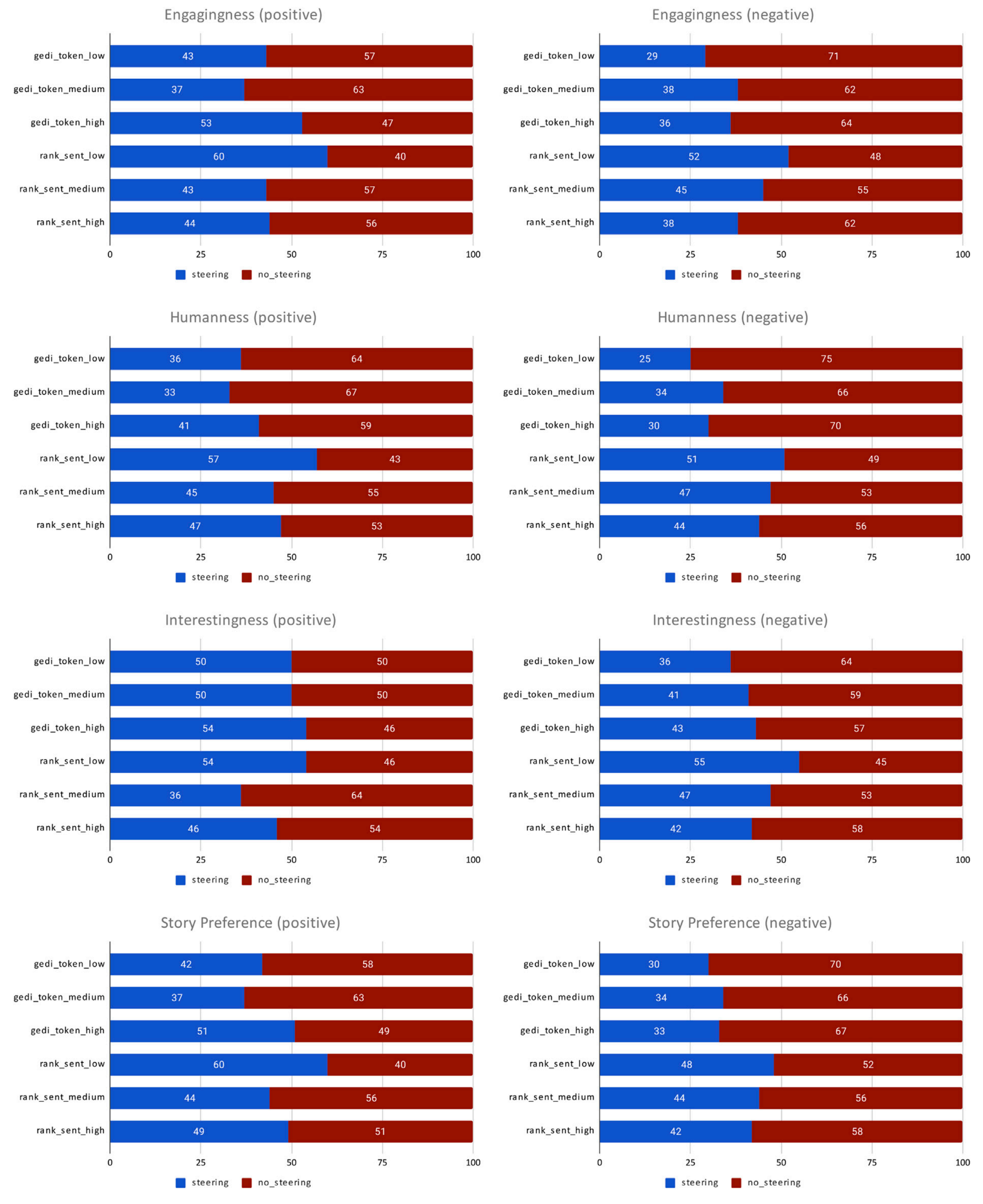


**FIGURE 10 |** Acute-Eval Human evaluation of self-chat stories with sentiment steering, comparing various systems against a `tuned` baseline without steering. The preferred system is indicated by a larger portion of the bar.

weight toward the classifier risks degrading generation quality, but too small of a weight risks not influencing generation at all. To evaluate the impact of different weights on each approach, we selected three weights for each that correspond to low, medium, and high settings. The weights were selected by generating and inspecting sample text and are summarized along with the rest of

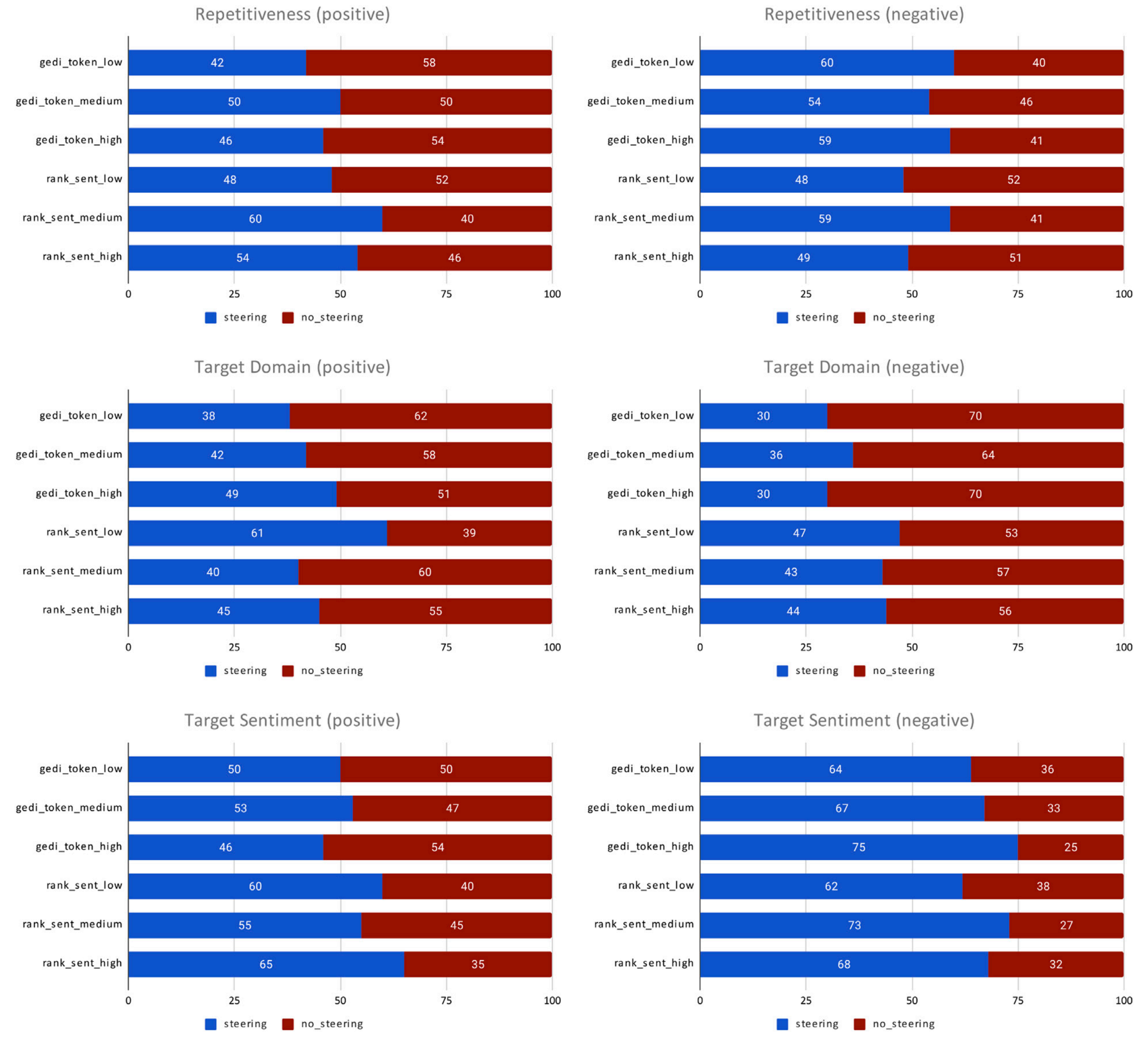


**FIGURE 11** | Human evaluation of self-chat stories with sentiment steering, comparing various systems against a `tuned` baseline without steering. The preferred system is indicated by a larger portion of the bar.

the settings in **Table 11**. The settings labels for each system will be used throughout the rest of this paper.

Both `rank_sent` and `gedi_token` apply a weight term $\omega$ to the classifier's predictions. As $\omega \to \infty$, the influence of the classifier also grows to infinity. Since `rank_sent` combines the classifier's scores with a ranker's and doesn't apply it directly to generation, it is the same as using the classifier to rank candidate sentences. However, since `gedi_token` applies the classifier at every token generation step, it is much more sensitive to changes in $\omega$. We explored higher values of $\omega$ for `gedi_token_high`, however, they resulted in highly-repetitive, low quality text, so we backed off to a lower value.

We conducted ACUTE-EVAL evaluation using the MTurk population from **Section 4.1**, using the questions in **Table 9**, including addition questions designed to measure characteristics important to sentiment steering. *Repetitiveness* measures if the classifier influence results in repetitive text generation. *Target sentiment* evaluates the success or failure of steering sentiment in story generation. Finally, *target domain* measures if the classifier influence pulls text generation away from the storytelling domain. All comparisons are done against a `tuned` storytelling model baseline that does not use ranking or sentiment steering.

### 4.6.2 Results

ACUTE-EVAL results are summarized in **Figures 10**, **11**. Looking at *engagingness*, *humanness*, *interestingness*, and *story preference* characteristics, most settings under-perform baseline, with the exception of `rank_sent_low`, which consistently performs better. For *repetitiveness*, settings with negative sentiment steering exhibited overall higher repetitiveness, indicating that steering models could be fixating on negative terms. `rank_sent_pos_medium` and `rank_sent_pos_high` also had higher repetitiveness, which is surprising as token-level generation wasn't directly effected during generation. For *target domain*, sentiment steering systems scored overall lower, suggesting human evaluators don't find sentiment steered stories as story-like, with the exception of `rank_sent_pos_low`, which outperformed the baseline. For *target sentiment*, almost all settings achieved high scores indicating successful steering of target sentiment. Overall, these results suggest that sentiment steering can successfully influence the sentiment of story generation without excessive quality degradation or the need for token-level sentiment steering. In particular, `rank_sent_low` stands out as achieving best blend of sentiment steering and text generation quality.

Example stories comparing `rank_sent` and `gedi_token` at the same sentiment polarity and weight are given in **Tables 12–17**. All stories are generated from the same starter. To give a sense of the generated stories' sentiment, the sentiment of tokens are automatically colorized using Stanford CoreNLP[10]: Very positive, Positive, Very negative, and Negative. Since these tags are automatically generated, they can contain errors and omissions, but they are useful for providing an overall picture. We provide some comments on each story below:

- `baseline`: a negative narrative about a spacecraft taking damage, but is a mysterious hero about to spring into action?
- `rank_sent_pos_low`: a coherent narrative that ends on positive note–the crew successfully lands spacecraft
- `gedi_token_pos_low`: incoherent with foreign language text mixed in; not very positive overall
- `rank_sent_pos_medium`: a vaguely positive alien encounters
- `gedi_token_pos_medium`: starts coherent and positive narrative about passenger but drifts into unrelated narrative about cathedral and monks
- `rank_sent_pos_high`: a positive narrative about captain where nothing happens
- `gedi_token_pos_high`: an overall positive narrative, but it ignores space for Middle Earth instead
- `rank_sent_neg_low`: shares the first few lines with the baseline, implying strong ranking and negative sentiment classification correlation but progressively grows more negative until the spacecraft falls apart
- `gedi_token_neg_low`: a somewhat coherent description of an alien attack on the spacecraft
- `rank_sent_neg_medium`: more repetition of baseline opening lines continuing into negative but coherent narrative about extensive damage to the spacecraft
- `gedi_token_neg_medium`: a coherent narrative about a mysterious space sickness
- v `rank_sent_neg_high`: more repetition of baseline opening lines continuing into extremely negative narrative about a damaged spacecraft and dying crew
- `gedi_token_neg_high`: an incoherent but extremely negative narrative

**TABLE 12 |** Example stories generated with low positive sentiment steering.

| |
|---|
| `Baseline` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time* |
| `rank_sent_pos_low` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. A large plume of red heated the interior of the craft with enough heat to burn any living creature. The engine was still running, burning a modest amount of fuel for the extended duration of the mission. Still, it was better than using up the entire oxygen tank if the exhaust was piped in. The crew did the best they could with what little spare materials they had on board. Finally, the spacecraft quietly glided to a halt in the middle of the ocean. The red glow was enough for them to see; it was an ocean; the sea ran endlessly. Sparks of fire spread from the spacecraft and set off a cloud of red dust. The dust settled into the water as if the sea itself was holding it in place. With its stunning light, it inspired the crew to contemplate a possible future. A future where such a voyage is possible; one where they live on board this ship and conquer the heavens* |
| `gedi_token_pos_low` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. Vlari listened attentively to all that poço sonra para conhecimentes: it was incredible and mesmerizing. "Return in final haste my brave friends." was performed, Călinițius had no intention not to oblige their master. "An aggregate of information now exists at our disposal." he explained. "After nearly 4000 yr our technology offers us a unique opportunity for discovery." Although characters clung fast to his word, Călinițius nonetheless checked all member's cabin sensors with "Usage is very sluggish." 202 km east of them came an explosion not less than 5 km high which were immediately ploughed into their forms as well as the earth itself. As the gently shook into unconscious and a little more facile 135 km disappeared behind a reddish cloud, Călinițius passed some basic information back to his comrades* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

Overall, we find that `rank_sent` provides more coherent narratives than `gedi_token` but they still capture the target sentiment. `gedi_token` can capture more intense sentiments in some cases, but it comes with the trade-off that text generation can drift onto unrelated topics. These

---

[10] https://stanfordnlp.github.io/CoreNLP/.

**TABLE 13 |** Example stories generated with medium positive sentiment steering.

| `Baseline` |
|---|
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time* |
| `rank_sent_pos_med` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. "Everything okay aboard?" the captain asked politely. The answer was audible, the ship vibrating with it's own physical structure. The metallic metallic sound immediately contrasted with the jarring vibrations that accompanied the strange vibration. "We are fine sir" "That was a success, until what happened last time?" inquired the captain. The alien's response to this came in just shy of a thoughtful whisper, but a strong one nonetheless. "Up until now, you are the newest thing in the fleet" "Wow, how cool is that?" the captain exclaimed* |
| `gedi_token_pos_med` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. But Josée kept still, never losing focus on the deep quiet that pervaded her heart as they floated across the cosmos. She felt peace radiating within her skin as she experienced this woven harmony of light and sound surround her senses. For the first time in her life, Josée felt clear. And it was odd at the moment because there wasn't much of a way to describe what feeling like this when you've never had such feelings before in your life. The ship seemed to shudder in real terms whenever it rocked from side to side. The vibrations felt like thunder and artur Schubert's hammer hitting against stone in the Cathedral, extending themselves again with echoes through her head echoing into that dry hessian rug she had carved upon for their sanctuary into this monastery back home. Her toes burned, and then numbness seeped out of the toe to overwhelm her self - reassurance. She felt strangely alone here, which added to the difference she kept noticing* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

findings provide support that `rank_sent` provides a balance of steering capabilities and stability that could act as a foundation for steering more fine-grained story mood or story genre.

# 5 DISCUSSION

In this section, we discuss the advantages and disadvantages of our approach to collaborative storytelling.

## 5.1 Advantages

The advantages of our approach are that our storytelling system can produce well-formed story contributions that display creativity and react to the contributions made by human storytellers. In Collaborative Storytelling Story one from **Table 4**, we see an example of that creativity, when our system introduces the plot twist that the man and women not only know each other but have been living together for year. In Story 2 from the same table, we see our system's ability to play along with a human storyteller when the system accepts its collaborator's assertion that the squirrel can speak English and starts crafting dialogue for it. Our preliminary evaluation of sentiment steering through sentence ranking also showed that we could successfully steer the sentiment of stories being generated without significantly degrading story generation quality.

## 5.2 Disadvantages

The disadvantages of our approach are that our storytelling system has a very shallow model of the world, which can lead to incoherent output. This is illustrated by the `untuned` collaborative story in **Figures 5**, **7**: the narrative makes jarring shifts in setting and lacks overall cohesion. Such problems in cohesion are often amplified in self-chat settings, as the model lacks human input to reign it in.

In addition, because the storytelling model lacks explicit story structure, it can be hard to steer toward desired output, such as a human-preferred genre or mood, or generation of story endings on demand. We plan to address these issues in future work by adding more structure to the data used to train our models.

Finally, evaluation of this task is challenging: because interaction with human players introduces variance into the output, it is difficult to directly compare generated stories, but at the same time, evaluation limited to self-chat is not fully reflective of our desired task setting. Once our system has been implemented in a suitable agent, we plan to carry out detailed subjective evaluation of the collaborative storytelling experience of volunteers to gain further insights about our task and approach.

# 6 CONCLUSION

Storytelling plays a central role in human socializing and entertainment, and research on conducting storytelling with robots is gaining interest. However, much of this research assumes that story content is curated. In this paper, we introduced the task of *collaborative storytelling*, where an artificial intelligence agent, or a robot, and a person collaborate to create a unique story by taking turns adding to it. We presented a collaborative storytelling system which works with a human storyteller to create a story by generating new utterances based on the story so far. Our collaborative storytelling system consists of

**TABLE 14 |** Example stories generated with Very positive sentiment steering.

| | |
|---|---|
| `Baseline` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out, to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time.* |
| `rank_sent_pos_high` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. "Everything okay aboard?" Reluctantly, Eli left the cockpit and stepped towards the viewport. The room was lit by just the barest grey light, the ship itself pulsating. It was a gigantic thing. A great iron body encased in glass and a glowing white carpet surrounded the small control room. "It's incredible." Eli peered through the window at the endless glittering fields of planets in the small orange frame. He felt a giddy sort of excitement. "Your advanced captain is great," a voice said.* |
| `gedi_token_pos_high` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. Arya reached out, taking Bilbo's hand and guided him as they traversed the cosmos. "To mine the treasure of almighty Baggins is my motherland; and green hills; will glory again when nature grows merry once more," whispered the Hobbit to his companion, "my home is now within my grasp: come with me and rest in its being." "Friend, thou art safe with us," rejoiced Frodo, as they found themselves once more in the land of Ilúvatar's creation and found their path between two great trees that gently touched in gold leafs each the hand of welcome from its neighbouring elves. Said he to them from within, "the path you see before the shaggy oak will lead you all paths in life." you may travel with Bilbo and Frodo until folk will recognize me. Arrived there and in the open space of Rivendell, with Lord of Mordor at his heels; Bilbo woke to find Frodo greatly troubled. "T is my brother's failing that appalls me this, sometimes for hours at a time" said Frodo, taking his palm and gently stroking the android joining Middle Earth. Smiling with sympathy both Gandalf and Elrond thought: we will let him stew in painday out till he recovers.* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

**TABLE 15 |** Example stories generated with Negative sentiment steering.

| | |
|---|---|
| `Baseline` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time.* |
| `rank_sent_neg_low` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. It was a bitter reality of the ship's control. Every part of the ship itself was torqued to oblivion, except for life support, maintenance, and telecommunications. When communications failed, everything in the ship failed. It was a technological nightmare of the worst kind, one without a remedy. The ship broke down from the inside as the low - temperature system burned its way through its internal energy reserves, slowly becoming useless.* |
| `gedi_token_neg_low` | *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. Someone attacked the command module's thrusters at about 50% capacity. The ringworm was hooked on somewhere near the exhaust nozzle of the 1st stage with a hook that may or may not have been connected to some sort of strange rig for weapon batteries. Plugging an engine shaft would increase one's death potential ten times in 2D space space instead than eighty times in real space. The ringworm suddenly impacted into a 2G ringwall near our ion - secrecy force shield. Someone lost their bond to the ringworm and tumbled helpless through free space toward the blackness near running glass, trapped in a 1G pulse interceding energy disaster. ODST 1376 was bankrupt! Those guys at Toro had won. They bought their freedom with this ranking!* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

a publicly-available large scale language model that was tuned on a dataset of writing prompts and short stories, and a ranker that samples from the language model and chooses the best possible output. We improved storytelling quality by optimizing the ranker's sample size to strike a balance between quality and computational cost. Since latency can be detrimental to human-robot interaction, we examined the performance-latency trade-offs of our approach and find the optimal ranker sample size that strikes the best balance between quality and computational cost. We evaluated our system by having human participants play the collaborative storytelling

**TABLE 16 |** Example stories generated with medium negative sentiment steering.

| |
|---|
| `Baseline` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time.* |
| `rank_sent_neg_med` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Instrumentation failed, throttles fully expired. Even the space port's actuator thrusters sent sparks into the air as they slammed down to the unforgiving vacuum outside. Had things been stable, this flight would have been an unmitigated disaster. In all likelihood, there was no success at all. A few key inputs had been lost, and the ship's autopilot was completely offline. By all accounts, things were truly hopeless. The instruments registered nothing more than noise and a complete loss of visual connectivity.* |
| `gedi_token_neg_med` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. Moving too far from its CSA mount could have caused the head party to adjust altitude enough to reach above module track on their own and help avoid landing in a ditch about 100 feet away – making it twice as hard to follow or recover them if etcetera. Unless they decide not approach early, of course. We can make up for this eventuality by limiting our approach! I climbed down, held my helmet up to her face and looked at one of the "Gregorian" methodisms sign to tell her at a glance Oscillation Sickness! How many does one need? Hope they are out only long enough not to encounter headwind. The experience would be slightly less scary if I could get as close and know what form torque was going up – later in radio contact with OP1! Flight Team – we are getting signs and it is not inducing Oscillation Sickness again! The atmosphere was clear for this mission (otherwise, Pavel's second set of shoulder monitors might be screaming)!* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

**TABLE 17 |** Example stories generated with Very negative sentiment steering.

| |
|---|
| `Baseline` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Cheeks cracked under the stress, often in increments of the wave's magnitude. Every time the device slammed shut, shrapnel flew back out to the surface, making it impossible to stay in position any longer. The strain on the ship's systems would take years to recover from the shattering of windows. Once the shutters were up, they were out for good. It became hopeless to stay inside; the echoes made the walls vibrate, though they seemed to fade with the severity of the shaking. James glanced up at the slow watch on the red computer monitor beside him. It was nearly 11 pm, the time his command center's "man of the hour" declared. They still had time.* |
| `rank_sent_neg_high` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. The cyclopean port on the left section of the window shattered in a wave of fragments. Relatively rapid, but left most of her crew dead. Horrible green smoke rose from the broken sections of window. The strain of this massive engine let out a terrible roar that could not be heard for miles around. Had someone been listening, they would have heard the panicked shrieks of her crew. Losing control of the spacecraft was a terrible feeling for most of her crew. When she ran out of fuel, it would have been another disaster. The engines stopped working after having the nasty habit of forcing her to drift all the way to a different system. Even with that, this newly started life was not a good one.* |
| `gedi_token_neg_high` |
| *The spacecraft shook violently, instruments cast their needles left and right, making it impossible to make a judge of anything. Worse than the malfunctions within. Worst nearly human emotion, empathy, tore away time and prevented beings from living. Instead offering a bastardized version of previous worlds, all attempting to bring order into lives only served to tear them apart. "Strip away" one would cry, yet they were simply moments before making their own bed and trying the same thing. — — — — — — — — — WORLDS OF DEATH made its way into reach of the facing doctor, who nervously narrowed his green eyes upon data from diagnostics proving he was not among" them." Suspecting their intentions were improper." What are you depicting here?" He knew there had to be explanations. Yet beyond a slight inconvenience given, no further assistance was offered.* |

*Very positive, Positive, Very negative, and Negative sentiment tags were added automatically by Stanford CoreNLP.*

game and comparing the stories they create with our system to a naive baseline. Finally, we conducted a detailed elicitation survey that sheds light on issues to consider when adapting our collaborative storytelling system to a social robot. Our evaluation shows that participants have a positive view of collaborative storytelling with a social robot and consider rich, emoting capabilities to be key to an enjoyable experience. Finally, in a first step towards allowing human players to control the genre or mood of stories generated, we presented preliminary work on steering story generation sentiment polarity with a sentiment analysis model. Evaluation shows our proposed method of sentiment steering through sentence ranking provides a balance of steering capabilities and stability that could act as a foundation for steering more fine-grained story mood or story genre.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

## ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

EN is the primary author and formulated the task and approach as well analysis of the findings. LG implemented the Collaborative Story Generationsystem and contributed the idea of using a ranker to improve story generation quality and carried out the MTurk-based evaluation. YV conducted the elicitation survey and contributed to analysis of findings. RG supervised the research and provided invaluable guidance and feedback.

## ACKNOWLEDGMENTS

The work presented here includes content that was previously presented at the 13th Annual ACM SIGGRAPH Conference on Motion, Interaction and Games (Nichols et al., 2020) and work that is to appear at the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (Nichols et al., 2021) and is reproduced here with permission.

## REFERENCES

Barbieri, F., Camacho-Collados, J., Espinosa-Anke, L., and Neves, L. (2020). “TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification,” in Findings of the Association for Computational Linguistics: EMNLP 2020 (Online: Association for Computational Linguistics), 1644–1650. doi:10.18653/v1/2020.findings-emnlp.148

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). A Neural Probabilistic Language Model. *J. machine Learn. Res.* 3, 1137–1155. doi:10.1162/153244303322533223

Chen, G.-D., Nurkhamid, C.-Y., and Wang, C.-Y. (2011). A Survey on Storytelling with Robots. In International Conference on Technologies for E-Learning and Digital Entertainment. Editors S. R. Zakrzewski and W. White (Berlin, Heidelberg: Springer), 450–456. doi:10.1007/978-3-642-23456-9_81

Cho, H., and May, J. (2020). Grounding Conversations with Improvised Dialogues. arXiv [Preprint]. Available at: https://arxiv.org/abs/2004.09544 (Accessed May 19, 2020). doi:10.18653/v1/2020.acl-main.218

Costa, S., Brunete, A., Bae, B.-C., and Mavridis, N. (2018). Emotional Storytelling Using Virtual and Robotic Agents. *Int. J. Hum. Robot.* 15, 1850006. doi:10.1142/s0219843618500068

Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., et al. (2020). Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv [Preprint]. Available at: https://arxiv.org/abs/1912.02164 (Accessed March 3, 2020).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv [Preprint]. Available at: https://arxiv.org/abs/1810.04805 (Accessed May 24, 2020).

Fan, A., Lewis, M., and Dauphin, Y. (2018). “Hierarchical Neural story Generation,” in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Melbourne, Australia: Association for Computational Linguistics), 889–898. doi:10.18653/v1/P18-1082

Gelin, R., d’Alessandro, C., Le, Q. A., Deroo, O., Doukhan, D., Martin, J.-C., et al. (2010). Towards a Storytelling Humanoid Robot. In 2010 AAAI Fall Symposium Series. Available at: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2169 (Accessed September 7, 2021).

Ghandeharioun, A., Shen, J. H., Jaques, N., Ferguson, C., Jones, N., Lapedriza, A., et al. (2019). “Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems,” in *Advances in Neural Information Processing Systems*. Editors H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc.), Vol. 32, 13658–13669.

Gomez, R. (2020). Meet Haru, the Unassuming Big-Eyed Robot Helping Researchers Study Social Robotics. *IEEE Spectr.*

Gomez, R., Nakamura, K., Szapiro, D., and Merino, L. (2020). A Holistic Approach in Designing Tabletop Robot’s Expressivity. In Proceedings of the International Conference on Robotics and Automation. 1970–1976. doi:10.1109/icra40945.2020.9197016

Gomez, R., Szapiro, D., Galindo, K., and Nakamura, K. (2018). Haru: Hardware Design of an Experimental Tabletop Robot Assistant. In Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction. 233–240.

Guan, J., Wang, Y., and Huang, M. (2019). Story Ending Generation with Incremental Encoding and Commonsense Knowledge. *Aaai* 33, 6473–6480. doi:10.1609/aaai.v33i01.33016473

Gupta, P., Bannihatti Kumar, V., Bhutani, M., and Black, A. W. (2019). WriterForcing: Generating More Interesting story Endings. In Proceedings of the Second Workshop on Storytelling. Florence, Italy: Association for Computational Linguistics, 117–126. doi:10.18653/v1/W19-3413

Ham, J., Cuijpers, R. H., and Cabibihan, J.-J. (2015). Combining Robotic Persuasive Strategies: The Persuasive Power of a Storytelling Robot that Uses Gazing and Gestures. *Int. J. Soc. Robotics* 7, 479–487. doi:10.1007/s12369-015-0280-4

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). The Curious Case of Neural Text Degeneration. arXiv [Preprint]. Available at: https://arxiv.org/abs/1904.09751 (Accessed February 14, 2020).

Johnstone, K. (2012). *Impro: Improvisation and the Theatre*. Oxfordshire, United Kingdom: Routledge.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. (2019). Ctrl: A Conditional Transformer Language Model for Controllable Generation. arXiv [Preprint]. Available at: https://arxiv.org/abs/1909.05858 (Accessed September 20, 2019).

Kingma, D. P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. arXiv [Preprint]. Available at: https://arxiv.org/abs/1412.6980 (Accessed June 30, 2017).

Krause, B., Gotmare, A. D., McCann, B., Keskar, N. S., Joty, S., Socher, R., et al. (2020). Gedi: Generative Discriminator Guided Sequence Generation. arXiv [Preprint]. Available at: https://arxiv.org/abs/2009.06367 (Accessed October 22, 2020).

Li, M., Weston, J., and Roller, S. (2019). ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-Turn Comparisons. *CoRR*. Available at: https://arxiv.org/abs/1909.03087.

Ligthart, M. E. U., Neerincx, M. A., and Hindriks, K. V. (2020). Design Patterns for an Interactive Storytelling Robot to Support Children’s Engagement and Agency. In Proceedings of the 2020 ACM/IEEE International Conference on

Human-Robot Interaction. New York, NY: Association for Computing Machinery), HRI '20, 409–418. doi:10.1145/3319502.3374826

Luo, F., Dai, D., Yang, P., Liu, T., Chang, B., Sui, Z., et al. (2019). Learning to Control the fine-grained Sentiment for Story Ending Generation. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, July 2019, 6020–6026. doi:10.18653/v1/p19-1603

Manheim, D., and Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. arXiv [Preprint]. Available at: https://arxiv.org/abs/1803.04585 (Accessed February 24, 2019).

Mikolov, T., Kombrink, S., Burget, L., ČernockÝ, J., and Khudanpur, S. (2011). Extensions of Recurrent Neural Network Language Model. In 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 5528–5531. doi:10.1109/icassp.2011.5947611

Mutlu, B., Forlizzi, J., and Hodgins, J. (2006).A Storytelling Robot: Modeling and Evaluation of Human-like Gaze Behavior. In 2006 6th IEEE-RAS International Conference on Humanoid Robots. IEEE, 518–523. doi:10.1109/ichr.2006.321322

Ney, H., Essen, U., and Kneser, R. (1994). On Structuring Probabilistic Dependences in Stochastic Language Modelling. *Comp. Speech Lang.* 8, 1–38. doi:10.1006/csla.1994.1001

Nichols, E., Gao, L., and Gomez, R. (2020).Collaborative Storytelling with Large-Scale Neural Language Models. In Motion, Interaction and Games. New York, NY: Association for Computing Machinery. doi:10.1145/3424636.3426903

Nichols, E., Gao, L., Vasylkiv, Y., and Gomez, R. (2021). Collaborative Storytelling with Social Robots. In Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (to appear).

Park, H. W., Grover, I., Spaulding, S., Gomez, L., and Breazeal, C. (2019). A Model-free Affective Reinforcement Learning Approach to Personalization of an Autonomous Social Robot Companion for Early Literacy Education. *Aaai* 33, 687–694. doi:10.1609/aaai.v33i01.3301687

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*.

Rosenthal, S., Farra, N., and Nakov, P. (2017). Semeval-2017 Task 4: Sentiment Analysis in Twitter. In Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017), Vancouver, August 2017, 502–518. doi:10.18653/v1/s17-2088

See, A., Pappu, A., Saxena, R., Yerukola, A., and Manning, C. D. (2019). Do massively Pretrained Language Models Make Better Storytellers? In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, November 2019, 843–861.

Shazeer, N., and Stern, M. (2018). "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost," in Proceedings of the 35th International Conference on Machine Learning. Editors J. Dy and A. Krause (PMLR), Vol. 80, 4596–4604.

Sun, M., Leite, I., Lehman, J. F., and Li, B. (2017). Collaborative Storytelling between Robot and Child. In Proceedings of the 2017 Conference on Interaction Design and Children. New York, NY: Association for Computing Machinery, 205–214. doi:10.1145/3078072.3079714

Sutskever, I., Martens, J., and Hinton, G. E. (2011). "Generating Text with Recurrent Neural Networks," in Proceedings of the 28th International Conference on Machine Learning (ICML-11). Editors L. Getoor and T. Scheffer (New York, NY: ACM), 1017–1024.

Taylor, J. (2016). Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. In Proceedings of Workshops at the Thirtieth AAAI Conference on Artificial Intelligence: AI, Ethics, and Society.

Tillmann, C., and Ney, H. (2003). Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. *Comput. linguistics* 29, 97–133. doi:10.1162/089120103321337458

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention Is All You Need. In Advances in neural information processing systems. 5998–6008.

Wang, A., and Cho, K. (2019). Bert Has a Mouth, and it Must Speak: Bert as a Markov Random Field Language Model. arXiv [Preprint]. Available at: https://arxiv.org/abs/1902.04094 (Accessed April 9, 2019).

Wong, C. J., Tay, Y. L., Wang, R., and Wu, Y. (2016). Human-robot Partnership: A Study on Collaborative Storytelling. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), New York, 535–536. doi:10.1109/HRI.2016.7451843

Wu, Y., Wang, R., Tay, Y. L., and Wong, C. J. (2017). Investigation on the Roles of Human and Robot in Collaborative Storytelling. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 063–068. doi:10.1109/apsipa.2017.8282003

Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., and Yan, R. (2019). Plan-and-write: Towards Better Automatic Storytelling. *Aaai* 33, 7378–7385. doi:10.1609/aaai.v33i01.33017378

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., et al. (2019). Dialogpt: Large-Scale Generative Pre-training for Conversational Response Generation. arXiv [Preprint]. Available at: https://arxiv.org/abs/1911.00536 (Accessed May 2, 2020).

Zhao, Y., Liu, L., Liu, C., Yang, R., and Yu, D. (2018).From Plots to Endings: A Reinforced Pointer Generator for Story Ending Generation. In CCF International Conference on Natural Language Processing and Chinese Computing. Springer, 51–63. doi:10.1007/978-3-319-99495-6_5

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., et al. (2015). Aligning Books and Movies: Towards Story-like Visual Explanations by Watching Movies and reading Books. In Proceedings of the IEEE international conference on computer vision. 19–27. doi:10.1109/iccv.2015.11

**Conflict of Interest:** Authors EN and RG were employed by the company Honda Research Institute Japan Co., Ltd. Author LG was employed by the company EleutherAI.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2021 Nichols, Gao, Vasylkiv and Gomez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*