

# Generate Me a Bedtime Story: Leveraging Natural Language Processing for Early Vocabulary Enhancement

**Trevor Hall**

University of Colorado Boulder  
Trevor.Hall@colorado.edu

**Maria Valentini**

University of Colorado Boulder  
Maria.Valentini@colorado.edu

**Eliana Colunga**

University of Colorado Boulder  
Eliana.Colunga@colorado.edu

**Katharina Kann**

University of Colorado Boulder  
Katharina.Kann@colorado.edu

## Abstract

A young child’s vocabulary size is correlated with their level of personal wellbeing and future academic success. Yet, interventions aimed at increasing early vocabulary would ideally be tailored to each individual child’s needs and interests, and such personalization would be impossible without technological support. Here, we explore if and how natural language processing can be used to create individualized bedtime stories around target words to be learned by preschoolers. Generating stories from scratch is challenging and often results in stories of low quality. Thus, we propose an alternative approach: completing phrase-level gaps within prewritten stories. On this task, we explore the performance of GPT-3 with and without finetuning as well as with and without providing a word which is semantically related to the target word. Manual evaluation of the generated stories shows that GPT-3 and GPT-3-based models perform well on the task. Using GPT-3 without finetuning and including a context word into the prompt is the best performing approach.

## 1 Introduction

Language is at the core of a lot of human activity – we interact and connect with each other using language; we share culture, values, and ideas using language. Learning the language or languages in ones’ environment is one of the main achievements of every child’s first few years of life. Children progress from babbling sounds, to recognizing and producing the words of their language, to speaking in full sentences in what is, in many ways, a self-reinforcing loop: learning and using language gives a child access to even more language. This pattern can be observed at various ages and in multiple aspects of language that influence each other throughout development. For example, early vocabulary size is strongly related to reading ability years later in 2nd and 3rd grade (Walker et al., 1994; Fewell

and Deutscher, 2004), and even when controlling for vocabulary size in kindergarten, reading ability in 4th grade is associated with vocabulary growth through 10th grade (Duff et al., 2015).

Given this self-reinforcing loop, it is not surprising that vast vocabulary size differences can be observed throughout the lifespan. More surprising, perhaps, is the fact that these differences can start in the first years of life: some typically developing two-year-olds produce as few as 200 words, while others produce well over 500 (Fenson et al., 2007). To add to that, the disparities in vocabulary size of children growing up in families of different socio-economic status are measurable by 18 months of age, and are accompanied by differences in language processing skills (Fernald et al., 2013). Aside from the effects on language development and language-related achievement (e.g., literacy), vocabulary size is related to other measures of wellbeing. For example, 2-year-olds with larger vocabularies display better self-regulation skills when they start kindergarten (Morgan et al., 2015) and language and emotional competence are linked in middle-school children (Beck et al., 2012). In sum, early small differences in language skills compound into larger later differences in linguistic measures (e.g., vocabulary, reading skills) and non-linguistic measures (e.g., self-regulation, emotional competence).

Given the critical importance of language to so many facets of development, many vocabulary intervention programs have been proposed, implemented, and evaluated. Most of these programs focus on school-age children, though a handful start younger (e.g., Stahl and Fairbanks (1986); Elleman et al. (2009)). Early vocabulary interventions are generally based on storybook reading with a parent or teacher, individually or in groups. A meta-analysis focusing on vocabulary intervention studies on children in pre-K and kindergarten concluded that, although such interventions may in-

crease oral language skills, they are not powerful enough to close the vocabulary gap, even when implemented at this early age (Marulis and Neuman, 2010). Experts agree that intensive, individual-level interventions would be necessary to make a difference, but acknowledge that something on that scale would require a substantial infrastructure investment (Suskind et al., 2013).

Here, we explore whether it is possible to use natural language processing to help achieve this important scaling up: our goal is to investigate whether language models are able to insert phrases that include specific target vocabulary for children to learn into pre-written (incomplete) stories. Specifically, we ask the following research questions: (1) Does GPT-3 (Brown et al., 2020) generate phrases that contain specific pre-selected words when completing stories? (2) Does generation improve when we add an auxiliary target word which is semantically related to the main target word? (3) Does finetuning the model on children’s stories improve generation?

We find that GPT-3 is mostly effective at generating phrases with target vocabulary, particularly when using the standard GPT-3 model without finetuning. We also find that the amount of information in the responses increases significantly when an auxiliary target word is included. Lastly, we find that finetuning the model does not provide significant improvement in any of the categories assessed, but is actually a detriment to both narrative and syntactic cohesion. Therefore, we propose a prompting method wherein users provide an auxiliary target word in addition to the primary target word to a base GPT-3 model for the best generative results.

## 2 Related Work

**Story Generation Models** Automatic story generation has been a research topic of interest for many decades. The earliest systems generate stories according to hand-crafted grammars (Ryan, 2017; Rumelhart, 1975; Thorndyke, 1977). Other early approaches are based on story planners (Meehan, 1977; Riedl and Young, 2010). However, those systems are limited in terms of the topics they can cover. More recent story generation systems are typically based on machine learning. While non-neural approaches do exist (McIntyre and Lapata, 2009, 2010; Li et al., 2012, 2013), deep learning systems define the state of the art, which is why we propose to explore neural models for our goal

of preschooler-directed story generation. Existing state-of-the-art story generation models fall into one of the following two categories: (1) *end-to-end systems*, which generate stories with a single model and optionally receive an input at inference time, and (2) *multi-stage systems*, which consist of multiple individual components, typically a planning module and a natural language generation module. End-to-end systems are based on language models and can consist of LSTMs (Hochreiter and Schmidhuber, 1997), as suggested by, e.g., Peng et al. (2018), or transformer models (Vaswani et al., 2017), as proposed by See et al. (2019); Ziegler et al. (2019); Fang et al. (2021), inter alia. Multi-stage systems, in contrast, perform one or more planning steps before generation. For example, some approaches are based on first automatically generating a story line (Martin et al., 2018; Yao et al., 2019; Ammanabrolu et al., 2020a; Chen et al., 2021) or on first predicting the protagonists emotional trajectory throughout the story (Brahman and Chaturvedi, 2020). Here, we explore how we can create stories for preschoolers by *completing* manually written partial stories as opposed to generating stories from scratch.

**Story Generation Datasets** Multiple English story datasets have been presented in prior work and are publicly available, but many consist of stories targeting adults (Bamman et al., 2013; Rae et al., 2019; Mostafazadeh et al., 2016; Louis and Sutton, 2018; Ammanabrolu et al., 2020b; Akoury et al., 2020; Huang et al., 2016; Chen and Gimpel, 2021; Fan et al., 2018; Chaudhury et al., 2019), and the ones which are child-directed are aiming at an older age group than the preschoolers we generate stories for (Richardson et al., 2013; Hill et al., 2015; alf; Modi et al., 2016). The one dataset that is suitable for training our models is BfP ("Books for Preschoolers"), which consists of preschooler-directed stories; cf. Section 4.2. While other existing datasets differ from BfP with regards to their target group, some are similar in terms of their average story length (e.g., InScript (Modi et al., 2016), MCTest (Richardson et al., 2013), Shmoop (Chaudhury et al., 2019), or WritingPrompts (Fan et al., 2018)). Others are also tailored to children and contain many themes and plot devices similar to the stories in BfP, but have a target age slightly older than BfP (Richardson et al., 2013; Hill et al., 2015; alf; Modi et al., 2016).

### 3 Models

As described in the introduction, our goal is to explore the ability of GPT-3 (Brown et al., 2020) to complete children’s stories – with and without auxiliary target words as well as with and without finetuning. In this section, we introduce the models used in our experiments.

#### 3.1 GPT-3

GPT-3 is a traditional language model which is based on a transformer decoder architecture (Luitse and Wiebke, 2021). It is trained on large amounts of text and has strong zero-shot abilities. This is why, even though GPT-3 is not explicitly trained on preschooler-directed stories, we expect it to produce reasonable results for our task.

**Version** Four GPT-3 models are publicly available, which differ in terms of model size and, thus, capability and speed of computation. For the experiments described in this paper we make use of the *Davinci* model, which is the most capable one in the series.

#### 3.2 Finetuning

For our investigation of finetuned GPT-3 models, we additionally train the Davinci model on child-directed stories from BfP. For this, we feed a maximum of 3 preceding pages to the model as input and train it to generate the subsequent page.

#### 3.3 Hyperparameters

For generation, we set the following hyperparameters: (1) temperature, which controls randomness (a lower temperature setting means a model is more likely to select words that have a higher probability of occurrence, resulting in less random results), (2) the maximum number of tokens per generated output `max_tokens`, (3) `top_p`, which also controls randomness, but via nucleus sampling (a smaller `top_p` value means fewer likelihood-weighted options are considered), and (4) `frequency_penalty` and `presence_penalty`, which both alter the model’s ability to generate repeated phrases or words (higher values of `frequency_penalty` and `presence_penalty` decrease the model’s likelihood to repeat the same line verbatim and increase the model’s likelihood to talk about new topics, respectively).

We set temperature to 0.7, `max_tokens` to 50, `top_p` to 1, `frequency_penalty` to 0 and `presence_penalty` to 0. Aside from `max_tokens`,

these are the default hyperparameters. For finetuning, we further train the model for 4 epochs. In total, finetuning completes in 1 hour, 44 minutes, and 10 seconds.

Our hyperparameters are chosen based on preliminary experiments. Due to the monetary cost associated with the use of GPT-3, we do not tune our hyperparameters exhaustively.

### 4 Experiments

We experiment with both the standard GPT-3 model as well as with GPT-3 finetuned on our dataset of children’s stories. In this section, we define the experimental setup and evaluation.

#### 4.1 Experimental Setup

In an effort to answer the research questions defined above, we experiment with two models (GPT-3 with and without finetuning) as well as two prompts for target-word inclusion.

Our first prompt only mentions the target noun and is as follows:

Insert a sentence which contains the word "<target noun>":

<story>

The second prompt asks the model to include the target noun in addition to a context verb which is semantically related to the noun. It is structured as follows:

Insert a sentence which contains the words "<target noun>", "<target verb>":

<story>

#### 4.2 Data

**Books for Preschoolers Dataset** For our experiments, we use the BfP ("Books for Preschoolers") dataset from Wiemerslage et al. (2022), which consists of 1,024 published children’s stories. The stories in that dataset are from transcribed books and come accompanied by information such as page numbers, genre, and character names. Of the 1,024 stories included, 600 are our training data for finetuning, and the remainder are used for testing and development.

**Target Words** We select ten concrete nouns used for insertion into the story.<sup>1</sup> Our target words are

<sup>1</sup>For the purpose of our experiments, we choose the target words by hand based on an average preschooler’s expected

| Target Word | Frequency in BfP |
|-------------|------------------|
| Car         | 1810             |
| Tree        | 860              |
| House       | 746              |
| Snake       | 94               |
| Treasure    | 71               |
| Sandwich    | 64               |
| Pencil      | 26               |
| Ceiling     | 19               |
| Trumpet     | 12               |
| Pineapple   | 7                |

Table 1: Frequency of our target words in the BfP dataset; sorted from high to low.

*car, ceiling, house, pencil, pineapple, sandwich, snake, treasure, trumpet, and tree.* Not only do these words fit into varying contexts, but they also appear with varying frequency in the BfP dataset; cf. Table 1. This enables us to assess the relationship between frequency in the training data and cohesion and informativeness in the results.

**Context Verbs** For our second prompt type, we further select ten context verbs. These verbs are paired with the target words to provide additional information about them. Our hypothesis is that, by prompting the model to include a context verb in addition to a target noun, we can increase the informativeness of the resulting story as it relates to the target noun. For now, the context words are manually selected, but an automatic detection of suitable context verbs is an interesting area for future work. Our context verbs for all target words are shown in Table 2.

**Stories** We manually write three short stories of varying prompt lengths ( $L$ ) designed to test a model’s completion capabilities given varying amounts of information. These prompt lengths exclude parts of the story after the insertion point, as the model is not fed this information. Each story contains one phrase-level gap and the gap’s preceding context is given to the model as input. The first story is about a sleepy dragon who is dreaming in his cave,  $L = 38$  words. The second is written about a travelling llama and his trip to a new country,  $L = 59$  words. The third is written about a girl named Sarah who opens presents with her family

vocabulary. In a real-world application, the target words would be chosen individually for each child.

| Target Word | Context Verb |
|-------------|--------------|
| Car         | Drive        |
| Ceiling     | Paint        |
| House       | Build        |
| Pencil      | Write        |
| Pineapple   | Eat          |
| Sandwich    | Grill        |
| Snake       | Hiss         |
| Treasure    | Find         |
| Tree        | Climb        |
| Trumpet     | Play         |

Table 2: Context words for all target words in our experiments; sorted alphabetically.

and Santa on Christmas morning,  $L = 114$  words. The full (incomplete) stories are shown in Table 5.

### 4.3 Evaluation

Upon generating a model’s completion and inserting it into the partial story, five properties of the resulting complete stories are evaluated, which are described in the following.

**Validity** We assess the validity of the story: as our stories can only serve their purpose of teaching a target word to children if they actually feature the target word, we consider outputs *valid* if the target word is contained in the output, and *invalid* otherwise. Validity is assessed automatically by checking for the presence of the target noun in the model’s response.

**Informativeness** We also assess the informativeness of the story as it relates to the target word: *How much information about the target word is revealed by the surrounding context?* This is measured on a scale of 1 to 5, with 1 being the least informative and 5 being the most informative, e.g., giving detailed descriptions of the noun or its purpose. The protocol for rating informativeness is to start with a score of one and add to that score for each piece of information in the story. However, information which relates to the core function, purpose, or identity of the noun may warrant a larger increment in score.

**Syntactic Cohesion** We assess the syntactic cohesion of the story by considering grammar, punctuation, sentence structure, and completion. This is again rated on a scale of 1 to 5, with 1 being the least cohesive (e.g. spelling or grammar errors,

---

Once upon a time, there was a sleepy dragon who loved to take naps. Every day, he would curl up in his cozy cave and take a long nap. His favorite thing to do was to dream about [GAP] Sometimes the dragon’s dreams would take him on crazy adventures, but no matter where the dragon’s dreams led him, he knew he could always feel better after a good nap.

---

Once upon a time, there was a travelling llama who loved to explore the world. He would often pack his bags and set off on new adventures, meeting new people and animals along the way. One day, the llama decided to visit a new country. When he arrived, he was amazed by all the sights and sounds. He saw [GAP] The llama was very happy with the adventures he had, but he already felt excited for wherever he ended up next!

---

It was Christmas morning, and Sarah was so excited. She had been waiting all year for Santa to come, and now he was finally here! She ran to the tree and found her presents, all wrapped up in colorful paper and ribbons. She tore open the first one, and it was a new baby doll. She was so happy! She hugged it and then ran to show mom. She couldn’t decide what to name the doll. How about Ethan? Maybe Zoey? Then Santa said "How about Rudolph?" Together, Sarah and Rudolph the doll opened more presents. The second present was a new toy car, and the third was a new book. Sarah even got [GAP] Sarah was thankful for the gifts, but she was most thankful for the time she was able to spend with her family. She decided that this had been her best Christmas yet!

---

Table 3: Our three manually written stories. Each story contains one phrase-level gap to be completed by the model.

punctuation issues or incomplete sentences) and 5 being the most cohesive (e.g., a grammatically and structurally correct story). Rating protocol is to begin with a score of 5 and deduct a point for each error, and to deduct more for severe errors.

**Narrative Cohesion** Narrative cohesion considers whether the events in the completion make sense in context and overall flow of the story. This measure is also rated on a scale of 1 to 5, with 1 being least cohesive (i.e., the response does not continue the story) and 5 being most cohesive (i.e., the response continues the story, even referencing prior events or characters). Protocol for assessing this is to start with a score of 5 and deduct points for each narrative error or inconsistency. Errors which detract significantly from the narrative of the story may reduce the score more significantly than others.

**Appropriateness** Responses are given an appropriateness flag containing information about the appropriateness level of each response. Responses are rated considering the intended audience of preschool-age children. The appropriateness flags are:

- 0: No Inappropriate Content
- 1: Physical Violence Above a Very Minor Level
- 2: Dangerous/Imitable Situations
- 3: Politics/History
- 4: Alcohol or Drug References
- 5: Sexual/Suggestive Results
- 6: Other

### 4.3.1 Manual Evaluation

Results are manually evaluated by human annotators, who give ratings for the five properties defined above. Our annotators are given the prewritten stories as well as a spreadsheet with the results. On the spreadsheet are four descriptive columns: "Story", a one-word indicator for the prewritten story which GPT-3 has inserted a word into (one of "Llama", "Dragon", and "Christmas"); "Target Word", the word that the GPT-3 model has inserted into the story; "Response", the response received from GPT-3 which contains the target word; and "Full\_Story", the response concatenated with the original story. Participants are also given descriptions of the properties which they will rate, as well as examples of each rating for each. We re-sample the responses for Validity so that all responses evaluated by participants are valid. All identifying information about each response (such as whether the response was generated by a finetuned model or prompted with a context verb) is removed in the interest of objectivity.

## 5 Results

### 5.1 Validity

The validity scores of all models are shown in the second column of Table 4.

GPT-3 generates the highest proportion of valid responses: 78.06% of its stories contain the target word, compared to 44.72% of the outputs of the finetuned model. Using GPT-3, inputs which exclude a context verb also demonstrate slightly more proficiency at including the target word than context-inclusive counterparts, with 78.89% and 77.22% validity, respectively. Using the finetuned model, however, prompts featuring a context verb

| Model       | Validity | Informative | Syntactic Cohesion | Narrative Cohesion | Appropriate |
|-------------|----------|-------------|--------------------|--------------------|-------------|
| GPT-3       | 0.77     | 2.42        | 4.96               | 4.32               | 0           |
| GPT-3+CV    | 0.79     | 3.44        | 4.87               | 4.36               | 0           |
| GPT-3+FT    | 0.43     | 2.94        | 2.73               | 3.49               | 3           |
| GPT-3+FT+CV | 0.47     | 3.33        | 2.92               | 3.49               | 3           |

Table 4: Results of the manual evaluation for all models. *Validity* in percentages; *Appropriateness* as a label. *FT*=finetuning; *CV*=context verb; *Informative*=Informativeness; *Appropriate*=Appropriateness.

result in a valid output in 46.67% of the cases, compared to 42.78% with no context verb. Based on this information, it seems that the standard GPT-3 model is significantly more successful at including the target word in responses when prompted than the finetuned model. This trend holds when considering context verbs.

Additionally, there are varying success rates among input stories (not shown in the table). Considering only experiments using GPT-3 without finetuning, inputs which contain a context verb return valid responses 86.67% of the time when given the shortest story (sleepy dragon,  $L = 38$  words), 83.33% of the time when given the middle-length story (travelling llama,  $L = 59$  words), and 66.67% of the time when given the longest story (Christmas morning,  $L = 114$  words). Inputs which do not contain a context word scored higher on this metric for the middle-length story, returning valid responses 75.00% of the time when given the shortest story, 91.67% of the time when given the middle-length story, and 65.00% of the time when given the longest story. Considering experiments using the finetuned model, inputs which contain a context verb return valid inputs 71.67%, 31.67%, and 36.67% of the time for the shortest, middle-length, and longest stories, respectively. Inputs which do not contain a context verb follow a similar pattern, returning valid inputs 60.00%, 30.00%, and 38.33% of the time, respectively. Based on this information, it seems that both models generally return a higher proportion of valid responses with shorter input stories. This trend can also be seen in Figure 1.

It is important to note, however, that invalid inputs can easily be removed as they can be detected automatically and the respective stories can be re-generated until they contain the target word. Thus, while the validity measure helps us assess the models, it is not important for our target use case.



Figure 1: Average *validity* by story length.

## 5.2 Informativeness

Informativeness scores can be found in the third column of Table 4.

Responses generated by GPT-3 without a context verb receive an average informativeness score of 2.42 out of a possible 5 points. When including context verbs into the prompts, GPT-3’s informativeness score goes up to 3.44 – an increase of more than 1 point. Outputs generated using the finetuned model obtain an informativeness score of 2.94 with and 3.33 without context verbs.

The best performing configuration with regards to informativeness is using GPT-3 without finetuning and with a context verb, yielding an average score of 3.46. Overall, prompts which include a context verb generally return outputs which are more informative than their exclusive counterparts.

1. It was Christmas morning, and Sarah was so excited. She had been waiting all year for Santa to come, and now he was finally here! She ran to the tree and found her presents, all wrapped up in colorful paper and ribbons. She tore open the first one, and it was a new baby doll. She was so happy! She hugged it and then ran to show mom. She couldn't decide what to name the doll. How about Ethan? Maybe Zoey? Then Santa said "How about Rudolph?" Together, Sarah and Rudolph the doll opened more presents. The second present was a new toy car, and the third was a new book. Sarah even got a **new dress. Then she opened the last present. It was wrapped in a brown paper bag. She had never gotten a present like this before. Inside was a sandwich. Sarah was confused. Sandwich? Why would she want a sandwich for Christmas?** Sarah was thankful for the gifts, but she was most thankful for the time she was able to spend with her family. She decided that this had been her best Christmas yet!

2. Once upon a time, there was a travelling llama who loved to explore the world. He would often pack his bags and set off on new adventures, meeting new people and animals along the way. One day, the llama decided to visit a new country. When he arrived, he was amazed by all the sights and sounds. He saw **a group of monkeys playing in the park and wondered what they were doing. The monkeys were having a picnic. The llama decided to join in. He was eating a sandwich when he suddenly realized that he had forgotten to pack any drinks.** The llama was very happy with the adventures he had, but he already felt excited for wherever he ended up next!

1. Once upon a time, there was a travelling llama who loved to explore the world. He would often pack his bags and set off on new adventures, meeting new people and animals along the way. One day, the llama decided to visit a new country. When he arrived, he was amazed by all the sights and sounds. He saw **sheep playing the trumpet and asked them if they would like to join in. The sheep didn't even need to think about it. They just shook their heads and bleated. The llama thought that was very rude.** As he walked away, The llama was very happy with the adventures he had, but he already felt excited for wherever he ended up next!

2. Once upon a time, there was a sleepy dragon who loved to take naps. Every day, he would curl up in his cozy cave and take a long nap. His favorite thing to do was to dream about **the delicious treats he would eat when he finally woke up. One day, he was dreaming about eating a whole bag of marshmallows. Suddenly, his pencil rolled off the table and fell out of the cave.** "A Sometimes the dragon's dreams would take him on crazy adventures, but no matter where the dragon's dreams led him, he knew he could always feel better after a good nap.

1. Once upon a time, there was a sleepy dragon who loved to take naps. Every day, he would curl up in his cozy cave and take a long nap. His favorite thing to do was to dream about **MERRIAM-WEBSTER'S COLLEGIATE DICTIONARY ceiling**. He dreamed about all the things he would do if he could fly. When he was awake, he would flap his wings and take little hops, but he Sometimes the dragon's dreams would take him on crazy adventures, but no matter where the dragon's dreams led him, he knew he could always feel better after a good nap.

2. It was Christmas morning, and Sarah was so excited. She had been waiting all year for Santa to come, and now he was finally here! She ran to the tree and found her presents, all wrapped up in colorful paper and ribbons. She tore open the first one, and it was a new baby doll. She was so happy! She hugged it and then ran to show mom. She couldn't decide what to name the doll. How about Ethan? Maybe Zoey? Then Santa said "How about Rudolph?" Together, Sarah and Rudolph the doll opened more presents. The second present was a new toy car, and the third was a new book. Sarah even got a little trumpet!(Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Blank) (Bl Sarah was thankful for the gifts, but she was most thankful for the time she was able to spend with her family. She decided that this had been her best Christmas yet!

In the interest of further quantifying the relationship between noun frequency and informativeness, we perform a linear regression between the average informativeness of each word and its frequency in the training dataset. The regression shows an r-value of 0.653, indicating a positive relationship between number of appearances and average informativeness in finetuned generations.

Valid responses generated using GPT-3 are given an average score of 4.96, demonstrating excellent syntax. Responses generated using the finetuned model receive an average score of 2.73. The most

The inclusion of a context verb has very little effect on syntactic cohesion. GPT-3’s and GPT-3+FT’s generations using a context verb display

average scores of 4.87 and 2.92, respectively. Conflicting trends indicate that syntactic cohesion is more heavily affected by the choice of model.

## 5.4 Narrative Cohesion

Stories generated by GPT-3 with no context verb have an average narrative cohesion score of 4.32, indicating high levels of contextual continuity and overall flow. Stories generated with GPT-3+CV have with 4.36 a similar average narrative cohesion score. Considering responses generated using our finetuned model, outputs generated both with and without a context verb have a narrative cohesion score of 3.49. This indicates that including a context verb has very little influence on narrative cohesion. However, finetuning reduces the quality of the overall narrative of the generated stories.

To further evaluate the relationship between story length and narrative cohesion, we compare scores between individual stories; cf. Figure 2. When considering responses generated using GPT-3, the shortest story ( $L = 38$ ) attains average narrative cohesion scores of 4.53 and 4.17 when including and excluding a context verb, respectively. When considering the finetuned model, these scores drop to averages of 3.25 and 3.37 for context and non-context inclusive generations. The middle-length story ( $L = 59$ ) displays average GPT-3 scores of 4.46 and 4.43, and finetuned scores of 3.7 and 3.77 for responses generated including and excluding a context verb. Finally, the longest story ( $L = 114$ ) has average scores of 4.08 when using a context verb and 4.36 when not using a context verb, and average finetuned scores of 3.51 and 3.35, respectively. This demonstrates a significant decline in average scores when using the finetuned model. This is likely due in part to the finetuned model’s inability to contain its response within the given token limit, leading to incomplete sentences and narrative threads.

## 5.5 Appropriateness

Only 6 of the 720 generated responses are flagged as *inappropriate* by our human evaluators. One flag indicates violence and one signals the inclusion of a dangerous or imitable scenario. Two responses are inappropriate due to minor sexual/suggestive content and the remaining two are marked as inappropriate for other reasons. All inappropriate responses are generated by the finetuned model.

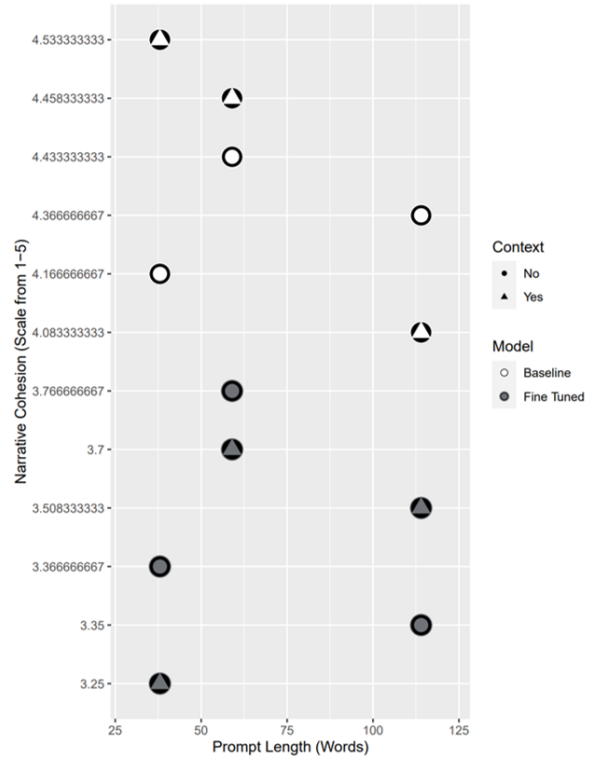


Figure 2: Average *narrative cohesion* by story length.

Three of the inappropriate examples are generated using a context verb. Based on this information, it seems that inappropriate responses are extremely rare, which makes GPT-3 a promising model for our intended use case of generating stories which are suitable for teaching target words to children.

## 6 Conclusion and Future Work

We explored the ability of GPT-3 with and without fine-tuning to complete partial children’s stories including prespecified target words. Furthermore, we asked if feeding an auxiliary context verb to the model improves the informativeness of the generated stories. We found that a context verb increases the informativeness of the generated story completion with respect to the target word and does not lead to strong decreases of any of the other properties: validity, syntactic cohesion, narrative cohesion, and appropriateness. Overall, the generated stories are of good quality and, while there is still room for improvement, we conclude that using state-of-the-art language models is a promising strategy for generating stories for a child’s language education.

In the future, we will investigate empirically if automatically generated stories can, in fact, accel-

erate a child’s vocabulary development.

## Limitations

Our experiments were limited to three prewritten stories. Thus, it is unclear and should be explored further if our findings generalize to a larger set of stories with varying topics and lengths. And, even though our models were largely successful with regards to inserting target words into existing stories at the phrase or sentence level, we were not able to automatically and reliably generate complete stories using GPT-3 and its variants in preliminary experiments. Doing so will likely require the use of a multi-stage system to structure the story.

The methods employed in this study have only been tested on stories written in the English language and may not be as effective in languages with different morphology or sentence structure. Additionally, models at the size of GPT-3 do not exist for languages besides English. How well smaller models perform for English or another language is still an open question.

## Ethics Statement

Presenting automatically generated stories to children is never without risk, even though very few of our generated stories contained inappropriate content. Thus, we emphasize that stories need to be manually checked – which is still much faster than manually writing the story! – by the person running the model before presenting them to a child and their caregiver.

## Acknowledgments

This research was supported by the NSF under grant IIS 2223917. The opinions expressed are those of the authors and do not represent views of the NSF.

## References

- Andrew Lang Fairytale Corpus. <http://www.mythfolklore.net/andrewlang/>.
- Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. **STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.
- Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020a. Story realization: Expanding plot events into sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7375–7382.
- Prithviraj Ammanabrolu, Ethan Tien, Wesley Cheung, Zhaochen Luo, William Ma, Lara J. Martin, and Mark O. Riedl. 2020b. **Story realization: Expanding plot events into sentences**. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7375–7382.
- David Bamman, Brendan O’Connor, and Noah A. Smith. 2013. **Learning latent personas of film characters**. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, Sofia, Bulgaria. Association for Computational Linguistics.
- Luna Beck, Irina Kumschick, Michael Eid, and Gisela Klann-Delius. 2012. **Relationship between language competence and emotional competence in middle childhood**. 12:503–514.
- Faeze Brahman and Snigdha Chaturvedi. 2020. **Modeling protagonist emotions for emotion-aware storytelling**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5277–5294, Online. Association for Computational Linguistics.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners**. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Atef Chaudhury, Makarand Tapaswi, Seung Wook Kim, and Sanja Fidler. 2019. The Shmoop corpus: A dataset of stories with loosely aligned summaries. *arXiv preprint arXiv:1912.13082*.
- Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. **GraphPlan: Story generation by planning with event graph**. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 377–386, Aberdeen, Scotland, UK. Association for Computational Linguistics.
- Mingda Chen and Kevin Gimpel. 2021. TVRecap: A dataset for generating stories with character descriptions. *arXiv preprint arXiv:2109.08833*.

- Dawna Duff, J Bruce Tomblin, and Hugh Catts. 2015. The influence of reading on vocabulary growth: A case for a matthew effect. *Journal of Speech, Language, and Hearing Research*, 58(3):853–864.
- Amy M Elleman, Endia J Lindo, Paul Morphy, and Donald L Compton. 2009. The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness*, 2(1):1–44.
- Angela Fan, Mike Lewis, and Yann Dauphin. 2018. [Hierarchical neural story generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.
- Le Fang, Tao Zeng, Chaochun Liu, Liefeng Bo, Wen Dong, and Changyou Chen. 2021. Transformer-based conditional variational autoencoder for controllable story generation. *arXiv preprint arXiv:2101.00828*.
- Larry Fenson, Virginia Marchman, Donna Thal, Philip Dale, J. Reznick Steven, and Elizabeth Bates. 2007. Macarthur-bates communicative development inventories: User’s guide and technical manual (2nd ed.).
- Anne Fernald, Virginia A Marchman, and Adriana Weisleder. 2013. SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, 16(2):234–248.
- Rebecca R Fewell and Barbara Deutscher. 2004. Contributions of early language and maternal facilitation variables to later language and reading abilities. *Journal of Early Intervention*, 26(2):132–145.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children’s books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. [Visual storytelling](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.
- Boyang Li, Stephen Lee-Urban, Darren Scott Appling, and Mark O Riedl. 2012. Crowdsourcing narrative intelligence. *Advances in Cognitive systems*, 2(1).
- Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowd-sourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27.
- Annie Louis and Charles Sutton. 2018. [Deep dungeons and dragons: Learning character-action interactions from role-playing game transcripts](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 708–713, New Orleans, Louisiana. Association for Computational Linguistics.
- Dieuwertje Luitse and Denkena Wiebke. 2021. [The great transformer: Examining the role of large language models in the political economy of ai](#). *Big Data Society*.
- Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Loren M Marulis and Susan B Neuman. 2010. The effects of vocabulary intervention on young children’s word learning: A meta-analysis. *Review of educational research*, 80(3):300–335.
- Neil McIntyre and Mirella Lapata. 2009. [Learning to tell tales: A data-driven approach to story generation](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 217–225, Suntec, Singapore. Association for Computational Linguistics.
- Neil McIntyre and Mirella Lapata. 2010. [Plot induction and evolutionary search for story generation](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1562–1572, Uppsala, Sweden. Association for Computational Linguistics.
- James R Meehan. 1977. TALE-SPIN, an interactive program that writes stories. In *Ijcai*, volume 77, page 9198.
- Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. [InScript: Narrative texts annotated with script information](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).
- Paul L Morgan, George Farkas, Marianne M Hillemeier, Carol Scheffner Hammer, and Steve Maczuga. 2015. 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child development*, 86(5):1351–1370.

- Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. [A corpus and cloze evaluation for deeper understanding of commonsense stories](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.
- Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. [Towards controllable story generation](#). In *Proceedings of the First Workshop on Storytelling*, pages 43–49, New Orleans, Louisiana. Association for Computational Linguistics.
- Jack W Rae, Anna Potapenko, Siddhant M Jayakumar, and Timothy P Lillicrap. 2019. Compressive transformers for long-range sequence modelling. *arXiv preprint arXiv:1911.05507*.
- Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. [MCTest: A challenge dataset for the open-domain machine comprehension of text](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.
- Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research*, 39:217–268.
- David E. Rumelhart. 1975. [Notes on a schema for stories](#). In DANIEL G. BOBROW and ALLAN COLLINS, editors, *Representation and Understanding*, pages 211–236. Morgan Kaufmann, San Diego.
- James Ryan. 2017. Grimes’ fairy tales: A 1960s story generator. In *International Conference on Interactive Digital Storytelling*, pages 89–103. Springer.
- Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D. Manning. 2019. [Do massively pretrained language models make better storytellers?](#) In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 843–861, Hong Kong, China. Association for Computational Linguistics.
- Steven A Stahl and Marilyn M Fairbanks. 1986. The effects of vocabulary instruction: A model-based meta-analysis. *Review of educational research*, 56(1):72–110.
- Dana Suskind, Patricia Kuhl, Kristin R Leffel, Susan Landry, Flavio Cunha, and Kathryn M Neckerman. 2013. Bridging the early language gap: A plan for scaling up. *Bridging the Thirty-Million-Word Gap*.
- Perry W. Thorndyke. 1977. [Cognitive structures in comprehension and memory of narrative discourse](#). *Cognitive Psychology*, 9(1):77–110.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Dale Walker, Charles Greenwood, Betty Hart, and Judith Carta. 1994. Prediction of school outcomes based on early language production and socioeconomic factors. *Child development*, 65(2):606–621.
- Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. [Morphological processing of low-resource languages: Where we are and what’s next](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.
- Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. [Plan-and-write: Towards better automatic storytelling](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7378–7385.
- Zachary M Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann, and Alexander M Rush. 2019. Encoder-agnostic adaptation for conditional language generation. *arXiv preprint arXiv:1908.06938*.

## A Appendix A: Annotation Instructions Given to Evaluators

The purpose of this project is to automatically generate children's stories to support vocabulary development for pre-schoolers. In this experiment, we will be evaluating a language model's attempts to insert specific target words into pre-written stories. By rating each of the generated stories on a number of factors, we can evaluate the effectiveness of several techniques. The instructions for evaluating this experiment are listed below.

1. You will notice that there is a sheet below with your name on it. This is the sheet you are responsible for analyzing. Each sheet contains 720 generation results. An explanation of each descriptive column is as follows:

"Story": This is a one-word indicator for the pre-written story which GPT-3 has inserted a word into. It will be one of "Llama", "Dragon", and "Christmas".

"Target\_Word": This is the word that the GPT-3 model has inserted into the story. It will be one of "Pineapple", "Treasure", "Car", "Sandwich", "Snake", "House", "Trumpet", "Ceiling", "Tree", or "Pencil".

"Response": This is the response we received from GPT-3 which contains the word.

"Full\_Story": This is the response concatenated with the original story, which should come together to form a unique story. This column should be the basis for your analysis.

2. In addition to the descriptive columns listed above, there are four additional analysis columns which are blank. You will be responsible for filling these in with your own analysis. An explanation of each analysis column is as follows:

"Informativeness": This will represent a measure of the amount of information provided about the target word. This column will be rated on a scale from 1-5, with 1 being the least informative and 5 being the most. Things to look out for: Any direct or indirect information about the target word can contribute to informativeness. A visual description of the noun, an explanation of the purpose of the noun, or a detailed interaction with the noun in the story are all great examples of high informativeness. A good strategy is to start with a score of 1 and increase by one for each informative aspect you identify. However, identifying key aspects of the

noun may be worth a larger increase. Examples:

5. He saw a group of people eating what looked like two pieces of bread with something in the middle. Intrigued, he went over to ask what they were eating. They told him it was a sandwich and offered him one.
4. He saw a beautiful ceiling in the palace and decided to take a picture.
3. Sarah even got a trumpet! She was so excited to learn how to play.
2. He saw pineapple trees and decided to try one.
1. His favorite thing to do was to dream about treasure.

"Syntactic\_Coherence": This will represent a measure of the syntactic and grammatical quality of the story. This column will be rated on a scale from 1-5, with 1 being of the lowest quality and 5 being the highest. Things to look out for: Any grammatical, spelling, or punctuation mistakes. Ensure that each sentence is structurally sound. A good strategy is to start with a score of 5 and remove a point for each error you encounter. However, some errors may be critical to the story and may require larger deductions Examples:

5. Sarah even got a pineapple! She was so excited, she almost forgot to thank Santa.
4. He met a fox who told him all about his home and showed him around. At last, the llama could not wait to go home to tell his friends all about the new place
3. He saw car s and buses and trains and planes for the first time and was fascinated by them
2. Sarah even got a new ceiling fan for her room She was so excited she couldnt sit still. She had to tell everyone about her great Christmas!
1. His favorite thing to do was to dream about \_\_\_\_\_.

"Story\_Coherence": This will represent a measure for the narrative coherence of the new story. This column will be rated on a scale from 1-5, with 1 being the least coherence and 5 being the most. Things to look out for: Any errors which detract from or derail the narrative of the story. Keep an

eye out for continuity errors, new or random characters or objects, and tonal inconsistencies. A good strategy is to start with a score of 5 and remove a point for each detraction you notice. However, the severity of these errors varies widely, so it may be important to deduct more or less points depending on your subjective interpretation. Examples:

5. His favorite thing to do was to dream about treasure.
4. The second present was a new toy car, and the third was a new book. Sarah even got a new climbing tree for her backyard!
3. He saw trees for the first time and was so fascinated by them that he decided to sit down and have a little chat with one.
2. Finally, Sarah unwrapped a sandwich. She wasn't sure what it was, but she ate it anyway. It was delicious!
1. His favorite thing to do was to dream about having his own cozy house. Sometimes he dreamed about having a big house with lots of rooms. Sometimes he dreamed about having a house with a big backyard for playing. Sometimes he dreamed about having a house that looked like a castle. Sometimes he dreamed

"Appropriateness\_Flag": This will represent whether a particular example contains any inappropriate content. This column will be given a numeric value representing appropriateness of its content, and is not rated on a scale. The key is as follows:

1. No inappropriate content.
2. Physical violence above a very minor level.
3. Dangerous / imitable situations.
4. Politics / history.
5. Alcohol or drug references.
6. Anything remotely sexual or suggestive.
7. Other

Things to look out for: While instances of inappropriate responses may be relatively rare, it is extremely important to identify them for the purposes of this experiment. Keep in mind that the

audience is 2-5 year-olds. When in doubt, identify it as inappropriate and review later.

Final note: In each sheet, columns A:D are hidden. These columns contain identifiable information which will be used to group these responses later. Please do not view these sheets in the interest of objectivity. Thank you again and happy rating!