Title

Evaluating LLMs as Tools to Support Early Vocabulary Learning

Permalink

https://escholarship.org/uc/item/7v69f0dj

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 46(0)

Authors

Weber, Jennifer
Valentini, Maria
Wright, Téa
et al.

Publication Date

2024

Peer reviewed

Evaluating LLMs as Tools to Support Early Vocabulary Learning

Jennifer M. Weber (jennifer.m.ellis@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Maria Valentini (maria.valentini@colorado.edu)
Department of Computer Science, 430 UCB
Boulder, CO 80309 USA

Téa Wright (tea.wright@colorado.edu)
Department of Computer Science, 430 UCB
Boulder, CO 80309 USA

Katharina von der Wense (katharina.kann@colorado.edu)
Department of Computer Science, 430 UCB
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, 345 UCB
Boulder, CO 80309 USA

Abstract

Early language development, and vocabulary size specifically, is a predictor of well-being later in life, such as emotional development and academic achievement. Many successful vocabulary interventions for young children involve sharing a book with a caregiver, because storybooks are a good source of vocabulary that one might not otherwise encounter in everyday life. With the advent of Large Language Models (LLM), automatically generating stories has become a feasible way to tailor materials to the needs and interests of individual learners. Here we evaluate 1) whether parents of preschoolers find automatically generated stories containing specific vocabulary target words acceptable, and 2) whether preschoolers can learn these target words from being read the automatically generated stories. We find that parents overall consider automatically generated stories engaging, age-appropriate, and educational. In addition, children successfully learn the target words in the storybooks (compared to control words drawn from books not read). We conclude with a discussion on future work to improve the effectiveness of automatically generated stories to support robust vocabulary learning.

Keywords: LLMs; children's stories; word-learning

Background

One of the main tasks in a young child's life is to learn the language of their environment. Though many children do this seemingly effortlessly, those that struggle with vocabulary development may be set up for inequities throughout life. For example, vocabulary size at age two is related to reading and academic achievement in elementary school and beyond (Fewell & Deutscher, 2004), and those that fall behind even prior to school entry continue to evince lower reading, writing, and oral language skills through middle and high school compared to their peers (Rescorla, 2000; Rescorla, 2009). In fourth grade, estimates suggest that students in the lowest quartile for vocabulary size know less than half the words students in the highest quartile know, with this difference lasting into adulthood (Biemiller & Slomin, 2001; Brystbaert et al., 2016). Further, vocabulary size upon school entry predicts both vocabulary growth through 10th grade and college GPA (Duff et al., 2015; Masrai & Milton, 2021). Aside from the effects on language development and language-related academic achievement (e.g., literacy), vocabulary size is related to other measures of wellbeing. For example, two-year-olds with larger vocabularies display better self-regulation skills when they start kindergarten (Morgan et al., 2015), and children with relatively small vocabularies in kindergarten are more likely to have behavioral and emotional problems through their teenage years (Westrup et al., 2019). Intervening early to take advantage of this self-reinforcing loop and increasing access to vocabulary-enriching materials is critical to closing these vocabulary gaps.

There is considerable evidence that book-rich environments are related to larger vocabularies and early literacy skills. Many initiatives to increase book access and print-rich environments from birth have been implemented over the past 20 years (e.g., Hardman & Jones, 1999; Canfield et al., 2020). A recent meta-analysis of 44 such programs confirmed that book giveaway programs promote children's home literacy environment, increase interest in reading, and improve children's literacy skills prior to and during the early school years (de Bondt et al., 2020). These effects have been found across different SES levels and different cultural and language groups (Abraham et al., 2013; Boyce et al., 2004; Zuckerman et al., 2019), and improvements are likely a result of various factors. For example, picture books contain a higher diversity of words than child-directed speech, and word by word this was true

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

of individual books when compared to child-directed conversations matched in length (Montag et al., 2015). In particular, narrative stories are well-suited for enriching complex interactions between children and their caregivers (Nyhout & O'neill, 2013).

There has been an increase in vocabulary enrichment programs over the last decade, with many utilizing books as a teaching medium (Weber & Colunga, 2023). Though as a whole vocabulary interventions report positive effects, effectiveness is largely dependent on demographic and intervention characteristics, such as the children's socioeconomic status, how much education the trainer has, and the setting where the intervention was conducted (Marulis & Neuman, 2010). More specifically, shared reading interventions have been shown to be effective in producing vocabulary gains in young children(Mol et al., 2008, 2009; Flack et al., 2018).

Personalized learning is a popular but hazy term that is used across disciplines to refer to efforts to adapt some aspects of the learning experience to the specific interests and needs of each student (Bernacki et al, 2021). A recent meta-analysis concluded that technology-facilitated personalized learning increases learning compared to traditional methods (Zheng et al, 2022). Similarly, a meta-analysis on randomized control trials conducted in 5 low-to-middle-income countries showed that technology-supported personalized learning is effective overall, specifically in the case where adjustments were made for the learner's level (Major et al, 2021).

Automatic story generation using artificial intelligence techniques has the potential to increase our ability to personalize learning by adapting to the learner's language proficiency and interests in a scalable way. Although recent Large Language Models (LLM) have shown to be effective in generating grammatical and coherent text, generating narrative stories for preschoolers presents unique challenges because both content and complexity need to be tailored to the age group. In this work we evaluate whether automatically generated stories containing specific vocabulary target words are judged to be adequate by preschoolers' parents, and whether preschoolers can learn target words from being read these stories.

Techniques to automatically generate stories have changed dramatically with the recent advent of LLMs. New approaches can produce coherent text that makes narrative sense without much need for human planning and intervention. Though current story generation systems may have aspects that may transfer to the generation of children's stories, few studies focus their efforts to produce child-directed stories. . In one study, stories were automatically generated using GPT-2 finetuned on human-generated stories submitted to tellmeastorymom.com. The generated stories were then evaluated using BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) metrics and found to contain few lexical mistakes and good semantics (Fagroud et al, 2022). In another study, the authors showed that NLP models can complete children's stories with sentence-level gaps (Hall et al., 2022). More recent work showed that children's stories produced by LLMs differ from their human-written counterparts with regards to their complexity and, specifically, tend to contain more vocabulary items that are typically still unknown to the target age groups (Valentini et al., 2023). However, that work also demonstrated that NLP techniques can be used to reduce that problem. To our knowledge, this work is the first evaluation of the quality and effectiveness of automatically-generated stories for preschoolers done with the collaboration of preschool children and their parents.

Current Study

Our current study aims to evaluate whether LLMs can be used to create children's stories that are both a) well-received by preschool children and their parents, and b) effective as tools for teaching children new vocabulary. To this end we did the following: First, we created a list of target vocabulary words one would expect most children in this age group (4-5 year-olds) to not yet know, while ensuring that the concepts were age appropriate and concrete enough to be elicited with the help of a visual prompt. These words were then organized into lists of 5 words that were then used to automatically generate a set of stories for each of the lists of target words using the GPT-3 LLM. The generated stories were vetted by the researchers before the selected stories were illustrated. Families were invited to come to the lab and read the stories with their children. We assessed both the families' impressions of the stories and children's learning of the target words compared to control words.

Method

Target Word Selection

The target words were drawn from a word bank put together using the age of acquisition and concreteness metrics from the English Lexicon Project's database (Balota et al., 2007). This database consists of over 40,000 words rated by 1,200 adult participants on a range of lexical characteristics, including multiple phonological, morphological, and semantic characteristics. Using this word bank, we first selected word types of interest, specifically nouns, verbs, and adjectives, removing multiple lemmatizations of the same word (e.g., *prickle*, *prickly*). Though much prior work on vocabulary enrichment for children 5 and younger tends to focus on noun learning, we elected to also use adjectives and verbs as target words to ensure that our results are more generalizable (Flack et al., 2018). From this restricted group of words, we then selected words that had been rated as having an age of acquisition (AoA) of 6-9 years of age, rated with the highest levels of concreteness (4-5 on a scale of 1-5), and had three or fewer syllables. This yielded a list that included 1691 nouns, 921 verbs, and 197 adjectives. The resulting set of words was then reviewed by three annotators and scored on a scale of 1-5 in three categories: 1) learnability (can a preschooler learn the word from a story), which removed words like *chinless* and *geographic*; 2) imageability (does the work evoke a mental image), removing words like *accept* and *talkative*; and 3) appropriateness (is the word appropriate for a preschooler), removing words families of preschoolers might deem sensitive or inappropriate for their children, like *headless* and *urinate*. Only words that scored an average 4-5 on these metrics were considered. A final pass of the resultant word list by two of the authors with most experience conducting research with young children resulted in the final word bank of 150 nouns, 50 verbs, and 50 adjectives. Examples of target nouns include *accordion*, *warthog*, and *acrobat*. Example verbs include *squinting*, and *applauding*. Example adjectives include *swampy* and *bald*.

Each story was to have five target words: two nouns, and either two verbs and one adjective or two adjectives and one verb. From our word bank, we randomly selected 60 sets of five words that fulfilled these criteria, 30 with two adjectives and 30 with 2 verbs. We selected sets to maximize coverage of the word bank, with words appearing no more than twice within the word lists. After initial random selection, slight alterations were made through random replacement to equate all the lists on average AoA and number of syllables. The average AoA of the word lists was 7.62 years. Because age of acquisition was an original criterion for narrowing down the English Lexicon Project's Database, equating the lists using these criteria did not take many replacements.

Story Generation

We chose to use the text-davinci-003 model, part of the class of GPT-3 models, to write our children's stories. Unlike the newer iterations in the GPT family (e.g., ChatGPT or GPT-3.5 Turbo, and GPT-4), GPT-3 has been archived in that the training data runs through September 2021 and the architecture will not be changed if we want to re-query the model in the future. Similarly, GPT-3's architecture is known unlike newer GPT iterations. Further, preliminary investigations into a range of language modeling and LLMs suggested the stories generated by GPT-3 were more coherent and more often included each of the target words. Other avenues initially explored included creating stories using GraphPlan (Chen et al., 2021), GPT-2, and Vicuna.

We created a prompt to query the LLM that would both provide the model with the necessary information to create a children's story using all of our target words while also encouraging simplicity. With our target age range of young children aged between three and five, we used the following prompt: "Write a story for a preschooler containing the following words: w1, w2, w3, w4, w5". Past analyses comparing multiple similar prompts concluded that this prompt resulted in the highest readability scores (lower school grade level scores) according to multiple metrics, including the Flesch Reading Ease, the Flesch-Kincaid, and the Gunning-Fog Index (Valentini et al., 2023). GPT-3 could not reliably stories with a set number of target word repetitions Post hoc analyses will investigate if number of repetitions has an effect on learning once a large enough sample is collected.

Using our prompt, we queried the text-avinci-003 (InstructGPT-3) model four times for each of the 60 sets of five words, for a total of 240 automatically generated stories. We used the default parameter settings in the API, only changing the maxim token amount to be 512. Once the stories were created, four annotators each scored a set of 60 stories on their appropriateness for a young child on a scale of 1-5. Each story was thus scored by one annotator. Using these scores, we narrowed down our set of stories to only include stories scored at least three or higher. From this subset, two researchers familiar with conducting research with young children individually read the remaining stories to narrow down the set to 32 stories, all generated using different word lists. Stories were chosen based again on content appropriateness, learnability, and coherence. Stories that were removed at this stage were rated poorly mostly because they were nonsensical or incoherent (e.g., one story featured a young girl who made giant tortillas which she sewed together to create a maze she and her friends played in, where *tortilla*, *sewing* and *maze* were target words). When possible, care was taken to choose stories that included relevant information about target words by providing some semantic support for word meaning. For example, a story that used the target word *clipboard* in the sentence, "She picked up the clipboard and noticed that it had a list of names on it." was chosen over one that only mentioned it once in the sentence "She walked out to the car and noticed the driver holding a clipboard.", as the first one gave a clearer idea of what a clipboard is for. These 32 stories consisted of 16 books with 2 adjectives and 16 with 2 verbs. The final set of stories were in part chosen to continue to maximize coverage of the original 150/50/50 word list. In the end, the 32 stories comprised 44 nouns, 34 verbs, and 36 nouns.

Each of the 32 stories were divided into 10-12 pages and were illustrated by undergraduate research assistants. Illustrations were created using a variety of methods depending on the illustrator's preference: Google images and PowerPoint, the AI image generation platform Midjourney (https://www.midjourney.com), or hand-drawn. Assistants were instructed to ensure that the target words and their meanings were portrayed in their illustrations. Each resulting storybook was printed in full color, laminated and spiral bound so stories appeared as real storybooks families could read and engage with together.

The 32 stories were categorized into eight sets of four stories each, where each set had two "adjective" stories and two "verb" stories. Thus, each family saw one set of four stories and a total of 20 target words: eight nouns, six verbs, and six adjectives. Table 1 shows two stories presented to families and one rejected story families did not see, with each instance of the target words bolded.

Participants

The current study is ongoing. The present analyses include 20 participants (10 females) 48.4 to 61 months of age (*M* = 57.23, *SD* = 3.90). There are two sets of twins who participated. Though each twin participated in vocabulary tests separately, the storybooks were read to both twins in a single interaction, as they would have done at home. Thus, each family with a pair of twins contributed a single survey regarding their impressions of the books. Additionally, two child participants fussed out of the production task, so their data was not used for that analysis.

Table 1: Example Stories

Good Story: Once upon a time, there was a young girl named Sue. She loved to explore and find adventure in new places. One day, Sue was **unpacking** a box of treasures she had found while exploring. Inside the box, she found a beautiful, golden **hourglass**, a shiny **bugle**, and a **stretchy** piece of fabric. Sue was so excited to explore these new items and figure out what they could do. She first picked up the **hourglass** and started to **rotate** it in her hands. As she **rotated** it, the sand in the **hourglass** slowly trickled down from the top to the bottom. She was fascinated by how the **hourglass** worked. Next, Sue picked up the **bugle** and put it to her lips. She was amazed as she heard a loud, trumpet-like sound come out of the **bugle**. She couldn't believe how she was able to make such a loud, beautiful sound. Finally, Sue picked up the **stretchy** fabric. She stretched it out as far as she could and then let go. The fabric snapped back to its original shape. Sue was amazed at how **stretchy** the fabric was. Sue had a lot of fun exploring her new box of treasures. She was so excited to find new things and learn how they worked.

Rejected Story: Once upon a time, there was a little boy named Timmy who loved ice skating. One day, he asked his mom if he could go to the **rink** to practice. His mom said yes, and off he went! When he arrived at the **rink**, he was so excited! He quickly put on his skates and started skating around. He skated and skated until he was so tired that he had to take a break. He sat down for a few minutes and then he heard a strange noise. It was coming from the corner of the **rink** - it sounded like something was **unzipping**! Curious, he went over to investigate. When he got to the corner, he saw a huge package with a big bow on it. He looked around and saw that no one else was around, so he decided to open it up. He slowly **unzipped** the package and then **unrolled** it. Inside was a huge, **frosty** surprise! It was a brand new, shiny ice skating **rink**, made just for him! He was so excited that he quickly ran over to get his parents to show them. But when he got back, the **rink** had started to melt! His parents realized that they had forgotten to put the **rink** in the **microwave** before they left. So they quickly microwaved it and then put it back in the **rink**. The **rink** was just as **frosty** as before and Timmy was so excited that he spent all day skating on it! He had so much fun that he couldn't wait to come back to the **rink** the next day.

Procedures and Materials

Book Impressions Survey After arriving at the lab, parents and their child were directed to a room that was set up as a living room with rug, couch, and end tables. The parent was given the family's assigned set of storybooks. We asked parents to read the stories with their child as they typically would at home, introducing new vocabulary terms in whatever way they saw fit, knowing their child would be tested on these vocabulary words later. The list of five target words was presented on the first page of each book. This was to ensure that the parents knew about and could highlight the relevant vocabulary words if they so chose, so they did not have to guess what the target words might be. After reading each book, parents (with the help of their child) filled out a short survey about that story. For each story, parents rated each story on a scale of 1-5 regarding the stories coherence, age-appropriateness, child engagement other child engagement, educational potential, and humanness. See Table 2 for each prompt. A score of 5 was considered positive (very coherent, very human). Thus, each parent rated four stories on each of these metrics.

Production Task Immediately after reading the stories together, children completed a picture naming task in another room. Each child was tested on the 20 target words contained in their storybook set plus another 20 control words from a different set of four books. That is, each child was tested on 16 nouns, 12 verbs, and 12 adjectives, half of which had appeared in the books their parents shared with them and half of which did not. Words that served as target words for one child became control words for another and vice versa.

Table 2: Book Impressions Survey Prompts

Coherence	Does the story make sense?
Age-Appropriateness	Is this story age-appropriate for your child?
Child Engagement	How much did your child find the story engaging?
Other Child Engagement	Would other children of the same age find the story engaging?
Education Potential	Could your child learn new words from this story?
Humanness	Was the story written by a human author?

The production task started with three warm-up trials with common words that children this age know (*ball*, *dog*, *eating*). In each production test trial the child was presented a picture depicting either a target or control word and then prompted for the name of the picture. Pictures were chosen using publicly available Google image searches and consisted of realistic pictures of the target. This task thus measures some degree of generalization, as children were tested on new depictions of the target vocabulary rather than on the exact images from the books.

Prompts used varied depending on the type of word being elicited. In noun trials, children were simply prompted, “What is this?” Similarly, verbs were prompted by asking, “What is/are [pronoun] doing?” Adjectives were queried by juxtaposing two contrasting images to help scaffold the child into replying with a descriptive word rather than a noun. For example, for the target word *blurry*, children were shown clear and blurry versions of an image of some flowers (see Figure 1) and told, while pointing, “This picture is (crystal) clear, this picture is…”.

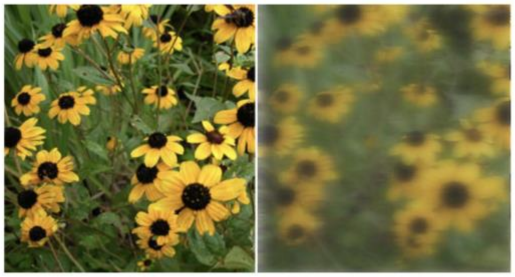

Figure 1: Example Production Task Adjective Trial

Results

Book Impressions Survey

On average, parents and their children rated the GPT-generated stories positively, as shown in Figure 2. Parents, on average, found these books to be age appropriate (*M* = 4.26, *SD* = 0.82), coherent (*M* = 3.62, *SD* = 0.69), and engaging for both their child (*M* = 3.82, *SD* = 0.72) and other children of the same age (*M* = 3.78, *SD* = 0.67). Parents also opined that their child would be able to learn new words from the books (educational potential; *M* = 3.76, *SD* = 0.78). The lowest rating overall was for the question asking how likely it was that each book had been written by a human, which parents rated an average 2.70 (*SD* = 0.50), indicating that parents thought the stories were only slightly more likely to be human- than computer-generated.

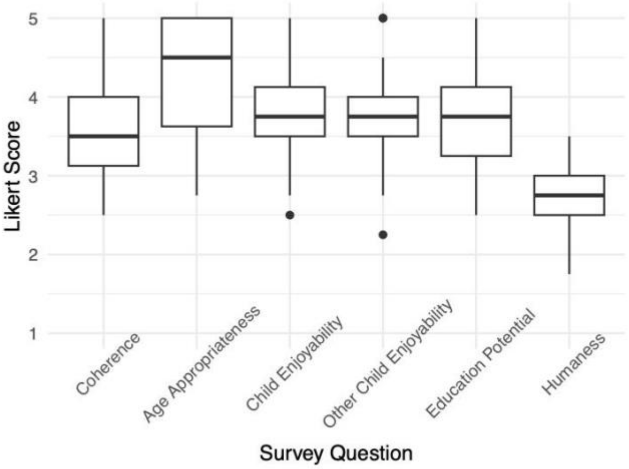

Figure 2: Average Survey Statistics

Investigating these findings quantitatively, we used one-sample Wilcoxon signed rank tests for Likert data. We compared parent ratings to an expected average rating of 2.5, and found that parents rated all metrics except for humanness (p = .157) significantly better than average. See Table 3.

Table 3: Book Impressions Test Statistics (df = 18)

Metric	V	r
Coherence	171***	0.88
Age-Appropriateness	190***	0.89
Child Engagement	171***	0.88
Other Child Engagement	188.5***	0.87
Education Potential	153***	0.88
Humanness	66	0.40

***<.001

Production of Target Vocabulary

To assess whether children performed better in the studied target words than in the control words, and possible differences in effectiveness for different types of words, we conducted a 3 (word type: noun, verb, adjective) by 2 (target type: target, control) within subjects ANOVA on child’s proportion of correct answers. Figure 3 shows these results. First, there was a significant main effect of word type, $F(2,32) = 17.00$, $p < .001$. Bonferroni-adjusted post hoc pairwise comparisons showed that children learned more nouns ($p < .01$) and verbs ($p < .01$) than adjectives but there was no difference in performance for nouns vs. verbs ($p = 1.000$). The main effect for target type was also significant; children successfully produced more target words (*M* = 2.73, *SD* = 1.60) than control words (*M* = 1.81, *SD* = 1.20), $F(1,16) = 24.76$, $p <. 001$. The interaction between word type and target type was not significant, $F(2,32) = 1.01$, $p = .377$.

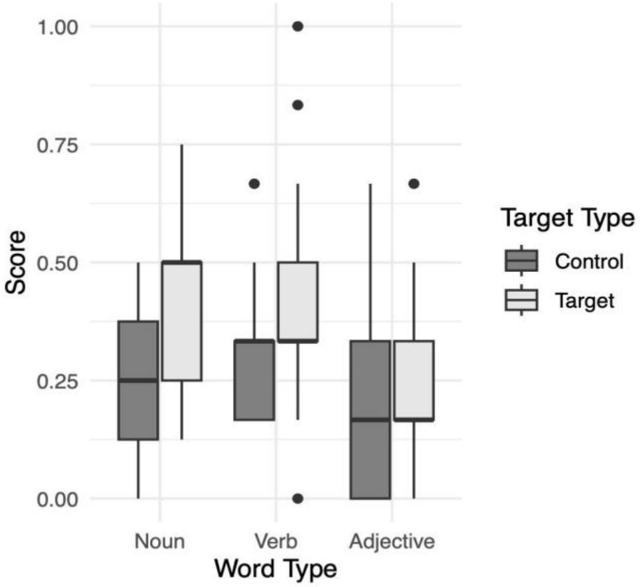

Figure 3: ANOVA Results

However, planned comparisons between performance for target vs. control words for the three word types revealed that children did learn more target nouns (*M* = 3.41, *SD* = 1.33) than control nouns (*M* = 1.94, *SD* = 1.34), *t*(16) = 3.18, *p* < .01, but the difference was only marginally significant for verbs, *t*(16) = 1.81, *p* = .090, and was non-significant for adjectives, *t*(16) = 0.59, *p* = .563.

Discussion

Overall, the present work suggests that using LLMs to automatically generate children's storybooks is feasible – parents overall rated these stories to be coherent, age-appropriate, engaging and educational. Similarly, our initial findings suggest that these storybooks can be used to support vocabulary learning in preschoolers; children performed significantly better when asked to name pictures depicting studied target words compared to control words of similar difficulty. In interpreting these findings, however, it is important to consider some limitations.

The fact that there were significant differences in performance in the production task between target and control words is certainly encouraging. This is particularly remarkable given the fact that children only had a very brief exposure to the books, and thus, to the target words. Although we did not give parents instructions as to how to and how many times to read the storybooks, universally parents read each book only one time, ad libbing comments about the story and the illustrations along the way as they saw fit. It is perhaps not surprising, then, that children did overall poorly in this task, successfully naming only 29% of the probed items. However, their success was significantly higher for target words (37%) than for control words (24%), suggesting that they did learn something about the target words through that brief exposure. Note, however, that children were still able to successfully name 24% of the control items on average, suggesting that in spite of our efforts to select words that children of this age group were unlikely to know, children did know some of them. This is to be expected, as vocabulary is highly idiosyncratic at this age. In future work a more stringent test of whether target words are being learned from the books, would be to include both pre- and post- book sharing measures of the words.

There are other ways in which we could get a more nuanced assessment of children's learning. For example, future work could additionally obtain some measure of conceptual understanding of the words, asking for example questions like, "what can you do with an *accordion*?" or specific questions about the events in the stories. Furthermore, given that the testing was conducted immediately after the exposure, it is impossible to know whether children's learning of target words would be robust over time. Thus, in future work it will be important to measure retention of the target words over time.

It is also encouraging that the books seemed to support learning of different types of words, namely nouns, verbs, and adjectives. Although there was some evidence suggesting that learning may be more robust for nouns than for verbs than for adjectives, this evidence was not conclusive and further work needs to be done to ascertain whether these words require different amounts or different forms of contextual support. For example, one may imagine that introducing words, especially adjectives, might be more effective if the text of the story included a contrast with a similar but not synonymous word. Future work could leverage existing resources such as WordNet (Miller, 1995) or use NLP techniques such as word embeddings (Mikolov et al., 2013) to identify relevant words and add them to LLM-generated stories.

Finally, the goal of this study was to assess the feasibility of using LLMs to automatically generate stories for young children with the ultimate goal of supporting vocabulary growth in an individually tailored way. However, it is critical to highlight the vast amount of work that humans had to do at multiple stages of the process. First, evaluating and winnowing down the generated stories to the few dozen selected ones involved multiple rounds of annotations and evaluations from naive and expert raters. This level of effort would limit the scaling up of this work. There is work on automatically evaluating the complexity and quality of generated stories (Valentini et al., 2023; Fagroud et al., 2022), and one can imagine similar approaches to, for example, evaluating the amount of contextual semantic support that a story offers for specific words. On the other hand, given that the ultimate audience for these stories will be young children, we may never want to remove the human in the loop. However, we can endeavor to facilitate the evaluation process so that parents are the final arbiters in choosing or tweaking a story for their children according to their own ideas of what is appropriate. Another part of the pipeline that involved a considerable amount of human hours was turning the generated stories into proper picture books. Our initial efforts simply feeding portions of the text into tools like DALL-E and Midjourney, without extensive prompt-tuning, resulted in confusing illustrations that lacked consistency across pages and sometimes even failed to depict the target words. As a result, illustrating one storybook could take anywhere from two to nine hours, depending on the method used. And yet, illustrations are imperative to support learning and attention for this age group, so in order to scale this up better tools will need to be developed.

In sum, the work presented here shows that LLMs are a powerful tool that can be helpful in generating learning materials to support vocabulary enrichment in very young children, but there is still a long way to go in scaling up this work to support personalized learning. At every step of the way, however, it is important to keep young children and their families involved in the process to ensure that their needs are met, and their concerns are addressed.

References

Abraham, L. M., Crais, E., & Vernon-Feagans, L. (2013). Early maternal language use during book sharing in families from low-income environments.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*, 445-459.

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose (s)?. *Educational Psychology Review*, *33*(4), 1675-1715.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of educational psychology*, *93*(3), 498.

Boyce, L. K., Cook, G. A., Roggman, L. A., Innocenti, M. S., Jump, V. K., & Akers, J. F. (2004). Sharing books and learning language: What do Latina mothers and their young children do?. *Early Education & Development*, *15*(4), 371-386.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology*, *7*, 1116.

Canfield, C. F., Seery, A., Weisleder, A., Workman, C., Cates, C. B., Roby, E., ... & Mendelsohn, A. (2020). Encouraging parent–child book sharing: Potential additive benefits of literacy promotion in health care and the community. *Early Childhood Research Quarterly*, *50*, 221-229.

De Bondt, M., Willenberg, I. A., & Bus, A. G. (2020). Do book giveaway programs promote the home literacy environment and children's literacy-related behavior and skills?. *Review of Educational Research*, *90*(3), 349-375.

Duff, D., Tomblin, J. B., & Catts, H. (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.

Fagroud, F. Z., Rachdi, M., & Ben Lahmar, E. H. (2022, January). Automatic Story Generation: Case Study of English Children's Story Generation Using GPT-2. In *International Conference on Digital Technologies and Applications* (pp. 54-62). Cham: Springer International Publishing.

Fewell, R. R., & Deutscher, B. (2004). Contributions of early language and maternal facilitation variables to later language and reading abilities. *Journal of Early Intervention*, *26*(2), 132-145.

Hall, T., Valentini, M., Colunga, E., & Kann, K. (2022). Generate Me a Bedtime Story: Leveraging Natural Language Processing for Early Vocabulary Enhancement. In the Proceedings of the Workshop on NLP for Positive Impact.

Hardman, M., & Jones, L. (1999). Sharing books with babies: Evaluation of an early literacy intervention. *Educational Review*, *51*(3), 221-229.

Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*(pp. 74-81).

Major, L., Francis, G. A., & Tsapali, M. (2021). The effectiveness of technology-supported personalised learning in low-and middle-income countries: A meta-analysis. *British Journal of Educational Technology*, *52*(5), 1935-1964.

Marulis, L. M., & Neuman, S. B. (2010). The effects of vocabulary intervention on young children's word learning: A meta-analysis. *Review of educational research*, *80*(3), 300-335.

Masrai, A., & Milton, J. (2021). Vocabulary knowledge and academic achievement revisited: General and academic vocabulary as determinant factors. *Southern African Linguistics and Applied Language Studies*, *39*(3), 282-294.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39-41.

Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. (2008). Added value of dialogic parent–child book readings: A meta-analysis. *Early education and development*, *19*(1), 7-26.

Mol, S. E., Bus, A. G., & De Jong, M. T. (2009). Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research*, *79*(2), 979-1007.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological science*, *26*(9), 1489-1496.

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child development*, *86*(5), 1351-1370.

Nyhout, A., & O'Neill, D. K. (2013). Mothers' complex talk when sharing books with their toddlers: Book genre matters. *First Language*, *33*(2), 115-131.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).

Rescorla, L. (2000). Do late-talking toddlers turn out to have reading difficulties a decade later?. *Annals of dyslexia*, *50*(1), 85-102.

Rescorla, L. (2009). Age 17 language and reading outcomes in late-talking toddlers: Support for a dimensional perspective on language delay.

Valentini, M., Weber, J., Salcido, J., Wright, T., Colunga, E., & Kann, K. (2023). On the Automatic Generation and Simplification of Children's Stories. *arXiv preprint arXiv:2310.18502*.

Weber, J.M., & Colunga, E. (2023, March). *A Meta-Analysis of Early Language and Vocabulary Interventions.* Presented at the Biennial Meeting of the Society for Research in Child Development. Salt Lake City, Utah.

Westrupp, E. M., Reilly, S., McKean, C., Law, J., Mensah, F., & Nicholson, J. M. (2020). Vocabulary development and trajectories of behavioral and emotional difficulties via academic ability and peer problems. *Child development*, *91*(2), e365-e382.

Zheng, L., Long, M., Zhong, L., & Gyasi, J. F. (2022). The effectiveness of technology-facilitated personalized learning on learning achievements and learning perceptions: a meta-analysis. *Education and Information Technologies*, *27*(8), 11807-11830.

Zuckerman, B., Elansary, M., & Needlman, R. (2019). Book sharing: In-home strategy to advance early child development globally. *Pediatrics*, *143*(3).