

Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling

Chao Zhang*
cz468@cornell.edu
Cornell University
Ithaca, NY, USA

Xuechen Liu
liuxuech@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Katherine Ziska
kziska@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Soobin Jeon
sbyeon@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Chi-Lin Yu
chilinyu@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Ying Xu
yxying@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

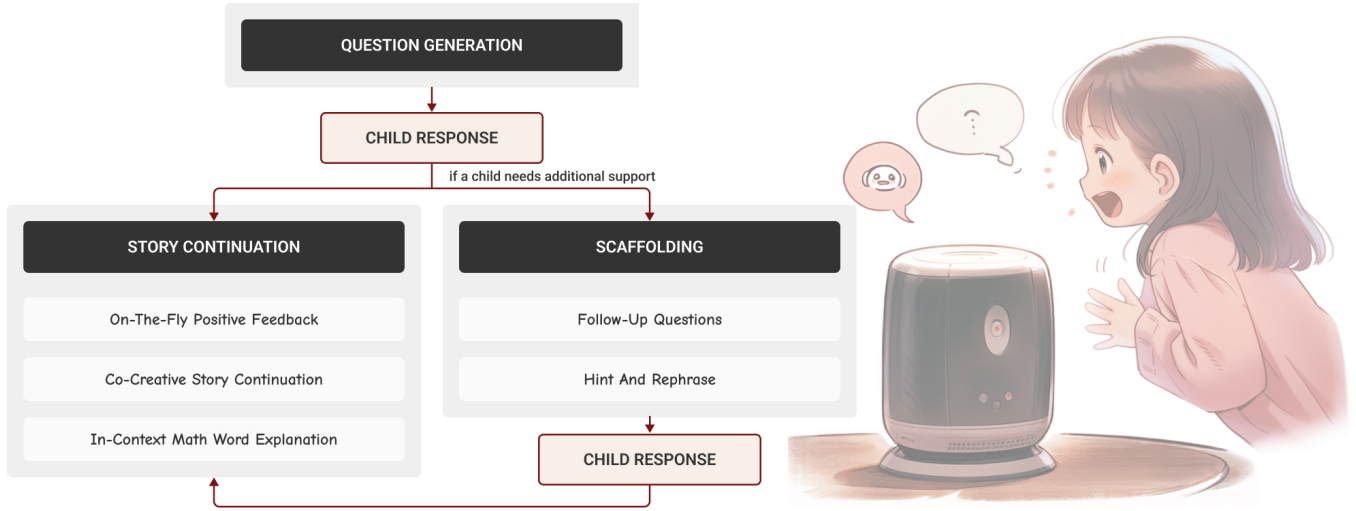


Figure 1: Interaction between a child and MATHEMYTHS: demonstrating the system’s ability to teach mathematical language through child-AI co-creative storytelling. MATHEMYTHS provides open-ended questions to solicit how the child wishes the story should progress, on-the-fly feedback to acknowledge the child’s responses, and co-creative story continuation with in-context explanations of math words. When the child needs additional support to continue the story, MATHEMYTHS offers scaffolding through follow-up questions and “hint & rephrase” strategies.

ABSTRACT

Mathematical language is a cornerstone of a child’s mathematical development, and children can effectively acquire this language through storytelling with a knowledgeable and engaging partner. In

this study, we leverage the recent advances in large language models to conduct free-form, creative conversations with children. Consequently, we developed MATHEMYTHS, a joint storytelling agent that takes turns co-creating stories with children while integrating mathematical terms into the evolving narrative. This paper details our development process, illustrating how prompt-engineering can optimize LLMs for educational contexts. Through a user study involving 35 children aged 4-8 years, our results suggest that when children interacted with MATHEMYTHS, their learning of mathematical language was comparable to those who co-created stories with a human partner. However, we observed differences in how children engaged with co-creation partners of different natures. Overall, we believe that LLM applications, like MATHEMYTHS, offer children a unique conversational experience pertaining to focused learning objectives.

*This work was carried out when the author was a research intern at the University of Michigan.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions.acm.org.

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0330-0/24/05...\$15.00

<https://doi.org/10.1145/3613904.3642647>

CCS CONCEPTS

• **Social and professional topics** → **Children**; • **Human-centered computing** → *Empirical studies in interaction design*; *Interactive systems and tools*; *Natural language interfaces*.

KEYWORDS

Storytelling, mathematical language, conversational interfaces, large language models, child-AI collaboration, co-creativity, children

ACM Reference Format:

Chao Zhang, Xuechen Liu, Katherine Ziska, Soobin Jeon, Chi-Lin Yu, and Ying Xu. 2024. Mathemyths: Leveraging Large Language Models to Teach Mathematical Language through Child-AI Co-Creative Storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 23 pages. <https://doi.org/10.1145/3613904.3642647>

1 INTRODUCTION

Mathematics is a fundamental building block in children’s early education and is a strong predictor of their future achievements [14, 20, 71, 82, 86]. Particularly in early childhood, mathematical learning is not just about learning to count or calculate; it also emphasizes the ability to understand and apply mathematical language, such as “half” or “equal”. Children often learn these words through everyday conversations with others [89]. For instance, when a parent asks, “Do you have more cookies than me?” children are introducing the quantitative term “more.” However, studies have found that the amount and quality of math language children encounter through these everyday, spontaneous conversations can vary greatly, which often aligns with socioeconomic differences [17]. This variation may have translated into disparities in early math abilities [25].

To enrich children’s experience of math language beyond everyday conversations, researchers have emphasized the importance of including a language component in general mathematics instruction, which allows children to engage with math language more systematically. However, it is challenging to teach math language abstractly to young children [60]. One of the approaches widely used with preschoolers and early elementary school children is to embed math language within the context of *narrative stories*, making abstract language more tangible and engaging for young learners [37, 76]. Typically, this story-based learning approach involves encouraging children to engage in dialogues within a storyline, where they understand and practice math terms with guidance from a more knowledgeable individual, such as a parent, teacher, or older peer [13, 21, 37, 65, 76]. Yet, the practicality of teaching math language through joint storytelling is debatable, as it demands a one-on-one engaged storytelling session that is logistically infeasible given the limited resources most educational institutions face [75–77].

In response to this challenge, researchers have long been investigating AI’s potential to simulate human-like storytelling interactions [16, 55, 69, 69, 84, 105]. However, before the emergence of Large Language Models (LLMs), AI was primarily optimized for structured interactions with children, such as asking questions and providing evaluative feedback [79, 102]. It is with the recent advancements in LLMs, notably GPT-4 [68], that the technical difficulty of facilitating spontaneous and creative conversations has

started to see promising developments [51]. However, the feasibility and efficacy of utilizing LLMs as partners to engage with children in storytelling remains unclear, especially in the context of supporting the learning of math language. Moreover, from a design perspective, it is unclear what specific workflow is necessary to tailor LLMs’ capacities to a younger audience [93].

This study outlines the development process for creating MATH-EMYTHS, a novel co-creative storytelling system powered by a GPT-based conversational agent, with the goal of teaching math language to children (Fig. 1). We then present empirical findings from a user study comparing learning and engagement outcomes using MATH-EMYTHS to those of a comparable human-guided interaction. Specifically, we aim to answer the following three research questions:

RQ1 *How can prompt engineering be employed to optimize LLMs as story co-creators tailored for children’s educational purposes?*

RQ2 *Do interactions with LLMs effectively engage children and provide an enjoyable experience during the co-creation of stories?*

RQ3 *Can children’s such interactions with LLMs result in comparable learning outcomes seen in human-child storytelling interactions?*

Our results indicate that when children interacted with MATH-EMYTHS, their learning of mathematical language appeared to be comparable to those who co-created stories with a human partner. However, while both younger and older children benefited from interacting with MATH-EMYTHS, we observed differences in the ways children from different age groups engage with co-creation partners of different natures. Overall, we believe that LLM applications, such as MATH-EMYTHS, can offer children a unique conversational experience, complementing their interactions with others. Our findings have implications for the design of LLM-based applications aimed at supporting children’s learning. Additionally, they highlight areas requiring further consideration to ensure the technology is both supportive and effective for diverse learning experiences.

2 RELATED WORK

In this section, we review research on (1) teaching math language through storytelling, (2) conversational interfaces for children, and (3) using LLMs for child-facing conversational interfaces.

2.1 Teaching Math Language through Storytelling

Children’s mathematical abilities are significantly influenced by their language skills [65, 71, 87, 96, 97, 100, 111]. Indeed, the Common Core State Standards, a US educational framework, highlight the need for students to be able to understand and apply specific mathematical language to describe the math concepts [3]. For instance, by second grade, students are expected to be familiar with terminology associated with fundamental math operations: addition, subtraction, multiplication, and division. Nevertheless, while some of this math language can be replaced by informal, everyday expressions (e.g., “putting together” as opposed to “addition”) to convey a similar meaning, equipping students with precise math terminologies is still necessary. Such terminology serves as a bridge to help children transition from their concrete everyday way of thinking to more abstract analytical thinking.

Children engage in math language in different ways, including their daily dialogues with teachers and parents [89]. Research has indicated a positive correlation between the extent and frequency of math language used by parents or teachers and the development of children's math abilities at the onset of elementary education [77]. However, the extent of math talk to which children are exposed varies significantly, likely influenced by the differing levels of awareness, expertise, or comfort among parents and teachers [17, 26]. This variation is recognized as a contributing factor to the disparities observed in children's math abilities [25]. Furthermore, while earlier studies predominantly focused on math language produced by parents and teachers, with children as recipients, more recent research has broadened this perspective. It emphasizes the opportunities for children to not only be exposed to but also generate math talk themselves [2, 83]. Interestingly, these studies have revealed that the children's own mathematical conversations play a crucial, if not more significant, role in shaping their mathematical competence, sometimes even outweighing the influence of the math talk provided by parents or teachers [90, 98].

To address the varying degrees of math talk children encounter in spontaneous circumstances and to encourage their active engagement in math discussions, researchers have developed various interventions. These interventions often utilize a storytelling approach [32, 36, 41, 46, 73], where children, along with their parents or teachers, engage in math-themed narratives. Discussion prompts are provided within these narratives to encourage children to use math language. For example, Purpura et al. [76] devised an eight-week intervention in Head Start classrooms, targeting children aged three to five. Researchers read six storybooks containing math vocabulary (e.g., "fewer") to the children. During the readings, researchers asked children questions that expanded upon the book's existing math language (e.g., "How do we know there were a lot and not just a few?"). Additionally, if a child expressed confusion about a math term, the researchers explicitly defined and explained it. The researchers found that children who participated in the intervention significantly outperformed those in a comparison group, who continued with business as usual, not only in a math language assessment but also in a math knowledge assessment. Another study used a similar story-based approach among kindergarten children with numeracy difficulties [36]. This study found that the students who engaged in the stories and related dialogue learned math vocabulary better than the other group of students who received direct instruction of the math vocabulary.

Traditionally, many story-based approaches have incorporated dialogue into existing narratives, thus, to some extent, constraining children's creative involvement in shaping the stories. Thus, the research community has also been promoting an approach that encourages children to contribute and decide how they want the narrative storyline to progress, which is believed to empower children to bring their lived experiences to the dialogue to support their meaning making [66]. This type of child-led storytelling has been adopted in preschool and early elementary classrooms and has proven to be feasible and age-appropriate. For instance, Flynn's study focused on a small group of four- to five-year-old children who were co-creating stories alongside an adult facilitator [28]. The research suggested that these children were able to meaningfully advance the story plot, drawing inspiration from ideas contributed

by their peers. It should be noted that, facilitation from an adult is important for children to fully engage in this process. In particular, the study highlighted the facilitator's role in posing questions to guide children's responses and offer constructive feedback responsive to children's contributions. Such *question-feedback-scaffolding* could potentially lower the cognitive demand required for certain challenges children face when continuing the story [102]. Building on this concept, the HCI community has long been exploring technology-mediated platforms to support children storytelling. We will discuss this line of prior research in detail in the section below.

2.2 Conversational Interfaces for Children

With the rapid advancement of AI, conversational user interfaces (CUIs) can now simulate interpersonal interactions through natural spoken language with children [72]. This type of speech-based interaction may not require children to have reading and typing skills, as is the case with other graphical user interfaces, thus removing the barrier for children who have not yet fully developed their literacy skills. Furthermore, it eliminates the need for a screen, potentially reducing tensions between children and their parents regarding screen time [39, 52]. Previous research has identified two prevalent types of interactions that children tend to have with CUIs [101]: voice assistants supporting open-domain conversations (e.g., Apple Siri, Amazon Alexa, Google Assistant) and voice-based apps that specifically designed for children for domain-specific conversations [8, 9, 56, 62, 63]. This type of specifically design apps have been found to support children's learning across a number of disciplines, including math [40, 98], science [102], computational thinking [18], and literacy [101]. Though varied in their specific design, these applications are grounded in a common principle by utilizing conversational AI as a language partner to engage children in discussions related to the relevant topics.

For instance, Ho and colleagues developed a social robot that provides math-related prompts in storybooks for parents of four- and five-year-olds, covering concepts like subtraction, addition, and comparisons [40]. The study revealed that parents found the robot's questions inspiring and that these questions led to more meaningful conversations with their children. Another example is a voice-based game designed to teach children aged four to seven [98]. In this game, children interacted with visual representations of shapes in different colors and sizes, receiving prompts such as "turn the tallest triangle into a circle." The game also featured back-and-forth questions to guide children through this process. A user study involving 18 children demonstrated that this game helped them learn the target math terms and engage in verbal reasoning about their decisions. Although these two studies utilized either the Wizard-of-Oz approach or pre-scripted dialogues, they provided insights and evidence of feasibility in terms of how CUIs could be designed to support math talk among preschool and early elementary-aged children. These insights have paved the way for our current research, which seeks to enable free-flowing, story-based interactions for children in the context of math dialogues.

Another related research area is focused CUI and storytelling specifically, where an agent collaboratively co-creates stories with children [97, 100, 103, 108, 109], mirroring the common childhood

activity children have with teachers or parents. Among these studies, a prevailing model involves the agent listening to a child's stories and then periodically offering generic, template-based response that are not usually responsive to children's specific contribution [10, 80, 94]. For instance, Bers and colleague developed a plush rabbit that reacts to a child's storytelling by moving its ears and shaking its body [11]. When the child pauses, the rabbit then prompts them with scripted follow-up questions to encourage further storytelling contributions [10]. This interactive strategy has been utilized in more recent studies as well [42].

While these studies have found that even this limited and structured interactivity can support children's storytelling, the agent made in fact minimal contributions to children's stories. Xu and colleagues referred to this type of interaction as "pseudo-conversation," where the agent follows predetermined dialogue flows when interacting with children [101]. This limitation could be especially problematic in creative storytelling activities, where children may generate a wide range of creative responses. Nevertheless, to overcome the technological limitations at the time, many studies employed Wizard-of-Oz approaches to explore the feasibility of children engaging in joint storytelling with a non-human partner, as perceived by the child. For example, Sun et al. [87] had children aged 4 to 10 interacted with a robot secretly controlled by a researcher, periodically inserting new story content and relating it to the story created by the child. Children were found to meaningfully collaborate with the "robot" and enjoy their interactions. These Wizard of Oz studies suggested that it is possible to replicate the strategies utilized by human partners in a setting with other technological partners, thereby laying the foundation for our research. With the advances in LLM, such human-led interactions can now be more feasibly simulated by AI agents. Thus, a primary objective of our current study is to utilize LLMs to emulate "true conversation," which is defined as multiple parties 'mutually orienting to, and collaborating in order to achieve, orderly and meaningful communication' [27]

2.3 Using Large Language Models for Child-facing Conversational Interfaces

Pre-trained large language models (LLMs), such as GPT-3 [12] and GPT-4 [68], have significantly advanced natural language processing (NLP) in recent years. These models, trained on colossal amounts of text data, can generalize to downstream tasks like text generation [93]. One key capability derived from the large model size is *prompting* [93], which allows individuals to provide specific textual instructions and examples [106] to guide the model's task execution. Although prompting LLMs may not consistently surpass benchmark models, it offers a lightweight approach to achieve competitive performance across a variety of tasks [12, 15]. The concept of prompting LLMs remains a compelling research topic within the HCI community. Our work extends previous research by introducing a set of prompting techniques iteratively designed to optimize LLMs as story co-creators tailored for children's educational purposes.

CUIs powered by LLMs, such as ChatGPT [67], have demonstrated their potential in facilitating responsive and engaging dialogues. Unlike traditional CUIs, they are not constrained by a

pre-determined dialogue paradigm or reliant on pre-written responses. One of their standout features is the ability to generate contextually relevant and coherent responses based on the input they receive so far [93]. This capability positions them to achieve "true conversation", wherein they can effectively collaborate with children, ensuring a more natural and enriching dialogue [56]. Thus, this paper harnesses the adaptability and flexibility of LLMs, aiming to forge a dynamic and collaborative storytelling experience for children, thereby fostering their math language growth and stimulating their creative expression.

Recently, numerous researchers have endeavored to incorporate LLMs into educational technologies [51]. These adaptations have seen the utilization of LLMs in creating educational content [19, 29], enhancing student engagement and interaction [1, 7, 88], and personalizing learning experiences [81]. For instance, LLMs have been used to generate children's narratives [4, 35], some of which have even been sold publicly [64]. In a different research trajectory, several scholars have used LLMs to create intelligent learning partners capable of collaborating with humans [47], providing feedback [49] and encouraging students [24, 88]. One common application involves employing LLMs as a conversational partner in written or oral form, such as in the context of task-oriented dialogues that offer language practice opportunities [24]. Building on these advancements in LLMs, our work aims to apply LLMs in CUIs to enable child-AI collaborative storytelling for mathematical language learning.

3 THE DEVELOPMENT PROCESS OF MATHEMYTHS

Grounded in the strong evidence that children develop math language through storytelling, and considering the unique capabilities of LLMs, we engaged in a design process to conceptualize MATHEMYTHS. Our aim was to create a co-creative storytelling system tailored for children aged 4-8, introducing them to mathematical language through voice-based interactions. This initiative serves as a case study to explore the potential of LLMs as educational partners for children.

At its core, MATHEMYTHS is designed to collaboratively engage in storytelling by alternating turns with a child. The role of MATHEMYTHS within this dynamic is three-fold: 1) generating prompts to elicit narrative contributions from the children, 2) continuing the story by utilizing mathematical language based on the children's input, and 3) providing scaffolding when children are hesitant to engage or show confusions about the prompts or the story. The interaction flow is presented in Fig. 2. The example conversation between a child and MATHEMYTHS can be found in Appendix A.

Given that using LLMs, particularly through prompt engineering, to develop educational tools is a relatively new domain, our design process began by applying educational principles from the existing literature reviewed above, which focus on engaging children in joint storytelling and math talk. We then engaged in a series of experiments to optimize MATHEMYTHS's capabilities for child-friendly and math-focused dialogue. In the remainder of this section, we outline 1) our overall design principles for MATHEMYTHS, 2) our method to prompt engineer LLM to serve as the story co-creator, 3) our process to evaluate the prompt-engineered LLM's performance, and 4) the implementation of the MATHEMYTHS system.

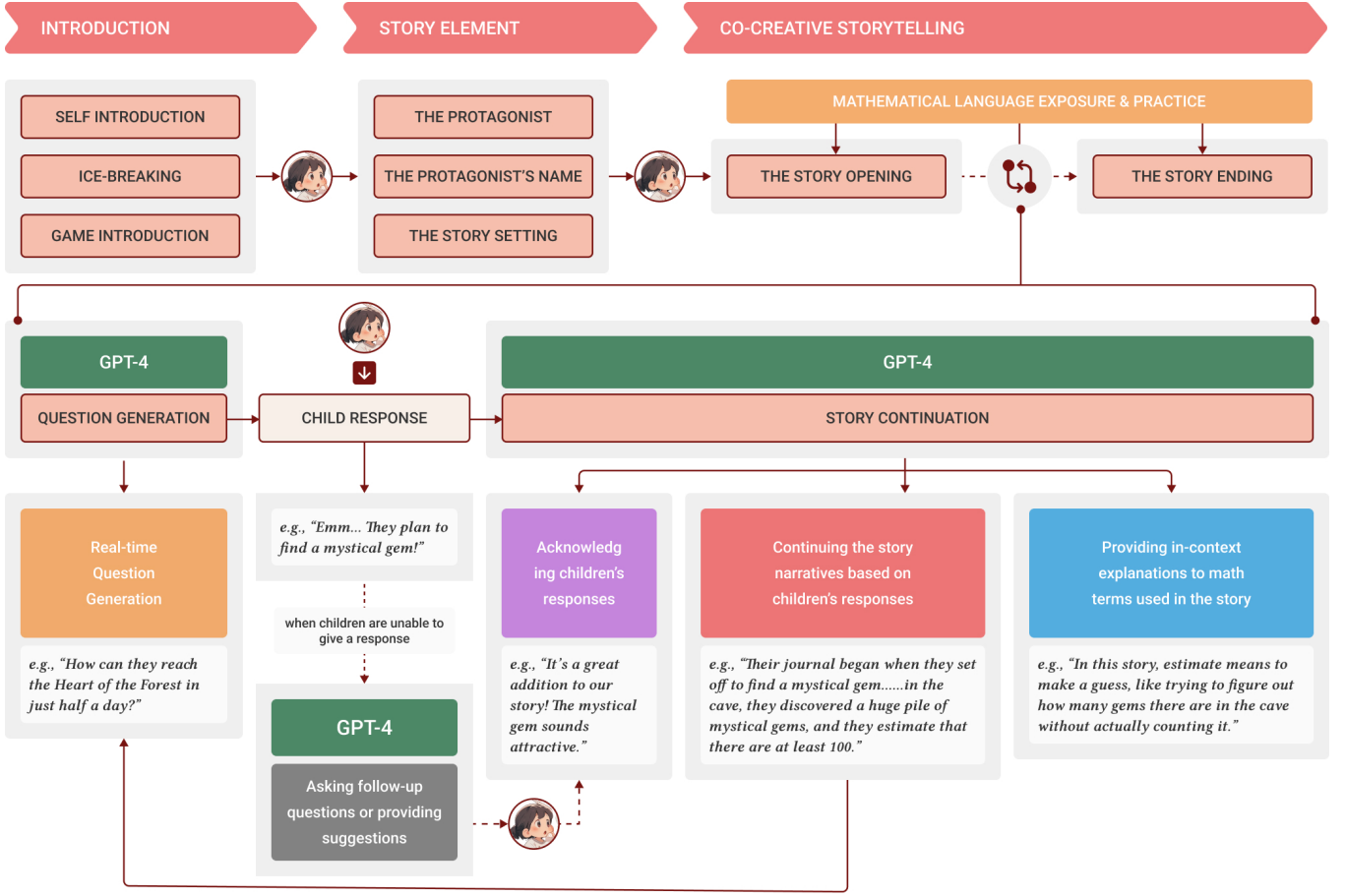


Figure 2: The interaction flow of the MATHEMYTHS system.

3.1 Design Principles

Drawing from the literature reviewed above, we incorporated two key design principles into the story co-creation interaction.

First, the conversation experience is designed to emulate a joint storytelling activity while the AI partner guides children in this process through using three-part dialogue moves that adhere to the *question-feedback-scaffolding* framework [102]. Under this framework, MATHEMYTHS initiates the interaction with children by crafting an initial story plot based on children's input about key elements such as the protagonist and setting. It then poses open-ended **questions** to solicit how children wish the story should progress, such as "What will Lucy and Nick do next to make sure they take an equal number of steps to reach the treasure?" After children's responses, MATHEMYTHS then offers **feedback** that includes both an acknowledgement of children's contribution (e.g., "Wonderful idea! That's a clever way to find the map.") and a continuation the narrative in line with the children's input. In the cases when young children require additional support to generate a concrete idea, MATHEMYTHS provides **scaffolding** to encourage children to elaborate their ideas [53, 87, 108, 109]. For instance, if the child provides a brief response without much detail, such as "ask for help", the agent will ask a follow-up question, "What a good point!

Who should we ask for help? Can you tell me more about these helpful friends?" In other cases, if the child indicates confusions (e.g., "I don't know") or does not respond, MATHEMYTHS will scaffold the child by employing the "hint and rephrase" strategy to provide ideas for children to consider (e.g., "Let's think together. Lucy and Nick can take turns counting their steps. How many steps do you think they should both count?"). This strategy, simplifying the original open-ended question by suggesting a specific direction, is widely utilized in conversational agents to engage preschool-aged children [33, 34, 102].

To make the storytelling experience more suitable for young children, MATHEMYTHS creates story based on the quest archetype. This genre, prevalent in children's literature, focuses on a hero's journey to a particular place or to find an item, facing many challenges along the way [44]. Language-wise, MATHEMYTHS is instructed to use simple, clear language that young children can easily understand, with sentences that are straightforward in structure.

Second, situated within the narrative structure, MATHEMYTHS is also designed to teach math language through exposure (i.e., using math terms in the plots generated by MATHEMYTHS) and practice (i.e., posing questions to encourage children to use math terms to

continue the stories). MATHEMYTHS primarily focuses on six mathematical terms (*sum*, *estimate*, *add*, *subtract*, *equal*, and *half*). These terms and their related concepts align with the kindergarten and first grade Common Core State Standards [3, 45]. The children in these grades, usually aged between 5 and 7 years, fall in the middle of our target age range. We also tested other terms from the Common Core Standard, such as “divide” and “multiply”, but ultimately decided not to use those terms as they might be too complex for our younger participants. Children heard these terms used at least twice per story. Moreover, MATHEMYTHS is uniquely designed to not only introduce these terms but also provide explanations within the narrative context. For instance, when the term “equal” is used in a sentence like, “Lucy and Nick take an equal number of steps,” MATHEMYTHS explains it in a story-relevant way: “This means Lucy and Nick take the same number of steps. If Lucy takes 5 steps, then Nick also takes 5 steps.” These types of explanations, integrated within the context, are deemed more effective for young children than abstract definitions [18, 32, 102].

3.2 Prompt Engineering

The model underlying MATHEMYTHS was GPT-4¹, currently the most advanced LLM available for developers. Through prompt engineering, GPT-4 performed three primary tasks aligned with the question-feedback-scaffolding model: 1) generating questions to facilitate children’s continuation of the story; 2) continuing the story based on children’s responses, including in-context explanations of math terms; and 3) providing scaffolding with additional suggestions or follow-up questions, tailored to children’s different responses.

The prompts were developed through an iterative process. We borrowed Brown et al. [12] suggested prompt engineering strategies to draft our initial prompts. We then underwent a two-month-long revision process, which involved internal evaluations with the research team and evaluations with children, and this process led us to rapid iterations of the prompts.

All these features were implemented using the OpenAI APIs². The final version of prompts used in the user study is displayed in Appendix D Table 4.

3.2.1 Question Generation. Following a similar prompt structure proposed in Brown et al.’s work [12], each prompt starts with a preamble which explains the prompt’s purpose: “*Given an unfinished story, compose one single, straightforward question to prompt a 6-year-old to expand on the story.*” The preamble is followed by a detailed list of instructions and multiple exemplars consisting of the input and the output. Specifically, we directed the model to pose questions concerning the main character’s subsequent actions or emotions. These two story elements are not only important to the narrative’s progression but are also tangible concepts that children can readily engage with and respond to [70]. To enhance GPT-4 performance, we provided six exemplar questions generated by the research team. These questions were further reviewed and approved by two trained research assistants in the domains of children’s verbal storytelling and mathematics education. The detailed prompt templates are shown in Appendix D Table 4.

3.2.2 Story Continuation. To facilitate the continuation of stories based on children’s inputs, while integrating mathematical language instruction, we began by prompting the model to assume the role of a “*storytelling robot*”³. This prompt established the interactive framework of the LLM’s responses.

The next part of the prompt was focused on context-setting. We supplied the model with a narrative scenario: “*Let’s play a joint storytelling game where we build a story together with the main character being a [character] named [name] in the [setting].*” The elements in brackets ([character], [name], [setting]) would be dynamically populated based on the responses provided by the children. We also set several high level rules to ensure the generated story content “*is simple and appropriate for a young child*”, adheres to the quest archetype, and contains nothing unrelated to the story.

For the model to produce coherent and math-focused responses, we articulated a specific sequence for it to follow. The model was tasked with acknowledging children’s contributions, utilizing a minimum of two mathematical terms, and then offering in-context explanations. This was achieved via the following detailed prompt: “*First, acknowledge my addition and commend me in a short 10-word sentence. Next, continue the story by correctly using [mathematical term one] and [mathematical term two] or their variants within two distinct short sentences. Lastly, explain their meanings within a short sentence, grounding it firmly within the story’s context.*”

During experimentation, we observed the LLM’s tendency to occasionally overlook prior directives as interactions continued. To counter this behavior and maintain consistency, our specific instructions were reiterated with each turn. This ensured MATHEMYTHS’s alignment with our intended conversational flow.

3.2.3 Scaffolding. In our study, we classified the situations where children require scaffolding into two main categories: (1) cases where they completely fail to continue the story, either by expressing uncertainty (e.g., saying “*I don’t know*”) or by not responding; (2) instances where they continue the story with only brief and vague responses, typically less than 5 words. To scaffold the first situation, our initial step is to determine whether the child is providing meta-comments (e.g., “*I like this*”, “*please continue*”) or exhibiting signs of confusion (e.g., “*I am not sure*”). This determination is made by prompting the LLM with a set of few-shot examples. These examples represent the most common types of responses observed in our rapid iteration tests with children. Following this identification, we utilize the model to simplify the question and offer hints related to the question, adhering to the “hint and rephrase” strategy. In addressing the second situation, we monitor the child’s responses. To scaffold the second situation, we employ the GPT-4 model to pose a follow-up question based on the child’s preceding brief response. This approach involves first acknowledging what the child has said and then encouraging them to elaborate further on their contributions.

3.3 Model Evaluation

As part of the iterative design process, we conducted an evaluation to examine the technical capabilities of our LLM-powered

¹<https://openai.com/gpt-4>

²<https://platform.openai.com/docs/api-reference/chat>

³In the prompt, the model is designated to function as a “storytelling robot”. However, during interactions, we avoid describing our system as a robot to prevent potential confusion, given that it operates as a virtual agent.

system. This evaluation consisted of two components: first, assessing the model's ability to *generate questions* to elicit children's responses to determine if these questions are fluent, inspirational, and logically relevant; second, evaluating the model's performance in *continuing stories* that are logically relevant, engaging, and age-appropriate. Specifically, our goal was to confirm whether the prompt-engineered GPT-4 model performs satisfactorily compared to a human baseline in both tasks and whether its performance surpasses that of a fine-tuned GPT-3.5 model in the question generation task. We chose the fine-tuned GPT-3.5 model as our baseline because it was the highest-performing model available that allowed fine-tuning during our development phase. Indeed, researchers have used a fine-tuned GPT-3.5 model in dialogic interactions with children, achieving better performance compared to other state-of-the-art models [56].

All evaluations were carried out by three college student research assistants who had over a year of experience in educational research and extensive years of engagement in activities with children within the target age range of our study. These students underwent training until their inter-rater reliability reached a satisfactory level before beginning the actual evaluation.

3.3.1 Evaluating Prompt Engineered GPT-4's Performance in Question Generation. The generated questions are designed to inspire children to continue the story and encourage them to learn and use mathematical terms. In this evaluation, we compared our prompt-engineered GPT-4 model with a fine-tuned GPT-3 model. The fine-tuned GPT-3 model was trained⁴ using a subset of 100 out of 300 human-crafted questions developed by an author of the paper. Besides these two LLMs, we also included the other subset of human-crafted questions as our baseline.

The question evaluation included four matrices, namely *readability*, *inspiration*, *story relevancy*, and *math language relevancy*. We invited the three evaluators to score the questions across each of the four dimensions in a 5-point Likert scale. The inter-rater reliability among these evaluators was satisfactory with an average of intra-class correlation of 0.82 across four dimensions. Details of these matrices are presented in Table 1.

An ANOVA was conducted to compare the ratings of questions generated by human, fine-tuned GPT-3.5, and our prompt-engineered GPT-4 model. The results indicated significant differences among the sources for all four dimensions (Readability: $F(2) = 9.415, p < .001^{***}$, Inspiration: $F(2) = 11.357, p < .001^{***}$, Story Relevancy: $F(2) = 17.327, p < .001^{***}$, Math Language Relevancy: $F(2) = 90.687, p < .001^{***}$). A post hoc Tukey's HSD test further revealed that questions generated by both humans and GPT-4 exhibited higher levels of readability, inspiration, and relevance to the story's plot and mathematical language when compared to questions generated by the fine-tuned GPT-3.5 model (Fig. 3). The evaluation ratings between human-generated and GPT-4 generated questions were comparable, further suggesting the feasibility and promise of GPT-4 models in this context. However, the questions, either generated by humans or GPT-4, were scored relatively lower in the relevance of mathematical language as compared to other dimensions but still higher than GPT-3.5. This arises from the complexity of creating questions that both meaningfully advance the

storyline and focus on specific math terms. To preserve the story's continuity, the questions' relevance to stories was often prioritized. Despite these, children should still encounter a sufficient number of questions aimed at enhancing their understanding and use of mathematical concepts.

3.3.2 Evaluating Prompt Engineered GPT-4's Performance in Story Continuation. MATHEMYTHS is designed to continue the story adaptively based on children's contributions and integrate mathematical language within the story plot. To evaluate whether our prompt-engineered GPT-4 model achieved these two goals, we compared the GPT-4 generated stories with those generated by humans as our baseline, given that other existing models were not tailored to continue stories but rather generate new stories. GPT-4 generated stories were created using a self-chat technique [30], which enabled our dialogue system to engage in self-generated conversations, mimicking a child's interactions, and collaboratively creating stories. Human-generated stories were based on the same instructions we provided to prompt GPT-4. The process for generating the dataset is detailed in Appendix B. Subsequently, we presented the three evaluators with pairs of stories, one generated by GPT-4 and the other by humans. Based on a widely recognized method for evaluating AI-generated dialogues [56, 58], we asked them to make a choice from each pair based on four dimensions: *preference*, *perceived creativity*, *mathematics relevancy*, and *readability*. Table 2 provides a detailed explanation of these metrics.

The results of this evaluation are displayed in Fig. 4. Overall, our evaluators indicated that the story pairs, where one was generated by humans and the other by GPT-4, were quite comparable in quality across all dimensions, particularly in perceived creativity (50.0% vs. 50.0%). Notably, there was a marginal preference for stories generated by AI (53.1% vs. 46.9%), with these stories also being perceived as integrating more mathematical elements (56.2% vs. 43.8%). GPT-4's advantages in relating stories to mathematical elements are likely due to the fact that this task capitalizes on LLMs' capabilities of creating seemingly logical connections of language elements based on specific instructions. However, human-generated stories were slightly favored for their readability (46.9% vs. 53.1%). A follow-up discussion with the evaluators revealed that the slightly lower readability in AI-generated stories primarily stemmed from instances where GPT-4 produced text that, while structurally coherent in its language, occasionally lacked logical consistency. This observation aligns with existing research underscoring the inherent limitations [48, 57].

3.4 System Implementation

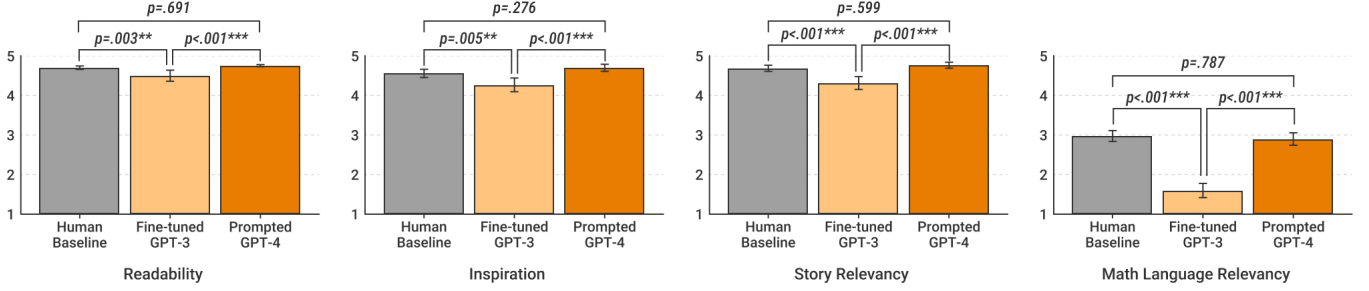
MATHEMYTHS is a desktop application designed to run on a MacBook computer with built-in microphones and speakers. This application, written in Python, does not rely on any graphical interface for interaction. Instead, users can navigate MATHEMYTHS simply by voice. Upon launch, users have the flexibility to specify both the number of dialogue rounds and the mathematical terms they wish to learn. For voice-based interaction, MATHEMYTHS leverages the Google Cloud Speech-to-Text API⁵ to recognize users' verbal

⁴<https://platform.openai.com/docs/guides/fine-tuning>

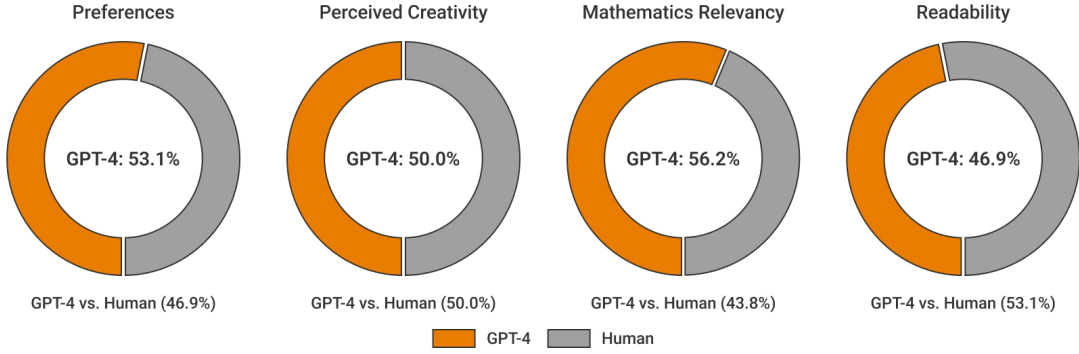
⁵<https://cloud.google.com/speech-to-text>

Table 1: Metrics used in evaluating prompt engineered GPT-4’s performance in question generation. The average intraclass correlation coefficient (ICC) scores are displayed.

Metric	Description	ICC
Readability	The generated question is written in understandable English, using proper grammar and vocabulary.	0.758
Inspiration	The generated question stimulates children’s thinking and provokes them to articulate their thoughts.	0.870
Story Relevancy	The generated question aligns with the narrative plotline.	0.751
Math Language Relevancy	The generated question requires children’s understanding or use of specific mathematical terms.	0.907

**Figure 3: Bar plots illustrating the distribution of data and the results from the ANOVA post-hoc Tukey’s HSD test regarding the question generation evaluation. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{**}$, $p < 0.001^{***}$. Error bars represent 95% confidence intervals (CIs).****Table 2: Metrics used in evaluating prompt engineered GPT-4’s performance in story continuation.**

Metric	Description
Preference	Who would you prefer to collaborate with to create a story?
Perceived Creativity	Whose story sounds more creative and contains more twists and turns?
Mathematics Relevancy	Whose story contains more mathematical elements?
Readability	Whose story is more easily comprehended by children in our target age range?

**Figure 4: The evaluation results on four metrics of story continuation.**

inputs. Additionally, the Elevenlabs Speech Synthesis service⁶ is employed to generate realistic speech for the agent’s responses.

4 USER STUDY

To understand how MATHEMYTHS might support children’s engagement and math language learning, we conducted a user study where children were randomly assigned to either co-create stories with

MATHEMYTHS or with a friendly and skilled human partner. We examined children’s gains in mathematical language knowledge, their engagement, creation performance, and enjoyment in the storytelling activities, and MATHEMYTHS’s performance.

4.1 Procedure

Participants were recruited from two public libraries in a mid-west city in the U.S. throughout the summer of 2023. This study was approved by the Institutional Review Board of the authors’ institution.

⁶<https://beta.elevenlabs.io/speech-synthesis>

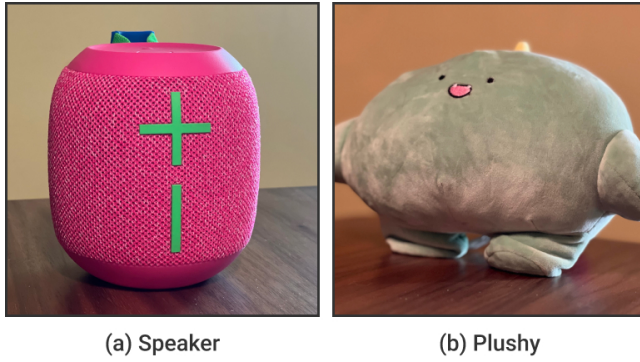


Figure 5: The (a) plushy and (b) speaker used in our user study.

After obtaining parental consent and child assent, participants were introduced to a research assistant and began the three-part study procedure which consisted of a baseline pretest of children’s math language, the story co-creation activity, and a post-test assessing math language, along with a survey with the children. Details of the math language pre- and post-test questionnaire are described in the Evaluation Metrics section below.

For the story co-creation activity, children were randomly assigned to partner either with a human (i.e., a train research assistant) or AI (i.e., MATHEMYTHS) presented as a colorful speaker as displayed in Fig. 5(a), for co-creating two short stories. In both conditions, children participated in a warm-up session to familiarize themselves with the story co-creation activity. This session also served to illustrate how the AI or human partner would precisely respond to their interactions, including scenarios where the children asked questions or chose to remain silent. To facilitate children’s story creation, they were handed a plushy (Fig. 5(b)) in case they wished to act out the story or just hold on to it. After the warm-up session, children were free to interact with the AI without interference or minimal redirection and instructions from the researcher. For both MATHEMYTHS and human conditions, the creation of each story typically lasted 6 minutes, resulting in a total duration of approximately 12 minutes for each child.

All parts of the procedure were video recorded for later transcription of the stories and coding of open-ended questions. Participants had the option of completing all three parts in a single session or completing the baseline assessment one day and returning another day to complete the story creation activity and learning evaluation. While children were completing the story creation activity, parents were provided with a questionnaire on an iPad to collect demographic information about their child. Once the child completed the procedure they were given a small prize and a book as a token of appreciation and parents received a twenty-dollar Visa gift card for their participation.

4.2 Evaluation Metrics

In this user study, our focus was on children’s learning of target math vocabulary, the quality of the stories they created, as well as their engagement and enjoyment throughout the process. For all measures requiring children’s responses, the questions were

orally narrated and asked by an experimenter, and the children were expected to answer orally as well. Thus, children did not need reading or writing skills to answer those questions.

4.2.1 Learning of Math Language. The research team developed a 24-item questionnaire focusing on the six math terms (i.e., *sum*, *estimate*, *add*, *subtract*, *equal*, *half*), based on the Common Core State Standards [3, 45] and Purpura et al. [74]. These terms were assessed through four dimensions: *definition*, *recall*, *transfer*, and *practice*, with each dimension comprising six questions, one question for each target term. For the *definition* dimension, children were prompted to provide the definition of each of the six math terms. In the *recall* dimension, they were asked to identify each term when its definition was given. For *transfer* questions, children needed to provide an appropriate math term based on a provided real-life scenario. Lastly, in the *practice* dimension, children were asked to use the target math terms to freely generate descriptions for a picture featuring fruits of various types, amounts, sizes, and colors. The complete questionnaire can be found in Appendix C.

During the pre and post-tests, we utilized the same set of questionnaires with slight modifications. For example, in the transfer set of questions, nouns were changed – (i.e., “you have a jar filled with candies” became “you have a net filled with butterflies”). Additionally, all items were presented in a random order. These modifications were implemented to reduce the likelihood that children were merely replicating their answers from the baseline assessment.

To calculate the scores from this questionnaire, the *definition* and *practice* items, being open-ended, were rated on a binary scale: 0 for incorrect and 1 for correct responses. For the *recall* and *transfer* items, children first attempted to answer through free recall. If unsuccessful, they were then provided with three options to choose from. Scores were assigned as follows: 2 for correct free recall answers, 1 for correct answers chosen from the given options, and 0 for incorrect answers. Based on this scoring system, we calculated a total score by summing the points across all items, with a possible range from 0 to 36. The Cronbach’s alpha of these items was 0.92.

4.2.2 Engagement. To understand children’s engagement in the story co-creation activity with either a human or AI partner, we analyzed their responses using three key indicators: whether they provided a verbal response, the length of each response, and the nature of their responses. The nature of responses was categorized into four types: meta-comments reflecting the child’s thought process (e.g., “I think so.”), expressions of uncertainty (e.g., “I don’t know”), brief responses with fewer than five words lacking sufficient details, and substantial responses containing meaningful details that advance the story. Two trained research assistants were responsible for the coding. The first coded transcripts for all participants, and the second coded 30% of the data for quality control purposes. Between these two coders, there was a 100% agreement rate across all items.

4.2.3 Story Creation Performance. We analyzed the children’s performance on their story creation using two indicators: the number of ideas and elaborations, based on prior research [96, 109]. An idea is defined as a character and its associated action within a single utterance. For example, “The cat climbs the tree” is one idea, with “cat” as the character and “climbs” as the action. An elaboration refers to additional details provided to enrich the idea, such as when,

where, and why. For instance, “The cat climbs the tree because it’s curious and loves to explore” adds reasoning for the character’s action. One researcher coded the transcripts for all participants, and another coded 30% of the participants’ transcripts as a quality check. This process resulted in a satisfactory Intraclass Correlation Coefficient of 0.85.

4.2.4 Enjoyment. To measure children’s enjoyment of interacting with AI or humans, we adapted a 4-item questionnaire based on Waytz et al.’s work [92]. The questions were: (1) *Were you happy when you told stories with []?*; (2) *Did you feel comfortable telling stories with []?*; (3) *Would you like to have a reading partner like []?*; and (4) *Do you want to tell another story with []?* The questionnaire asked children to indicate their level of agreement with each item, using a pictorial scale to facilitate their responses. For example, in response to the first question, children were presented with five options of “really happy,” “happy,” “kind of happy,” “a little bit happy,” and “not happy at all.” Each option was associated with a circle varying in size from large to small, with the researcher pointing to each circle while describing the corresponding option. The size of the circle corresponded to the level of happiness, with the largest circle representing the most happiness and the smallest indicating the least. To ensure that children based their responses on MATHEMYTHS and not on other voice assistants they may have used at home, we reminded them at the beginning that we were specifically asking about the AI they had just created a story with, while also pointing at the device. The Cronbach’s alpha internal consistency is acceptable at 0.77.

4.3 Participants

We recruited 35 participants (19 female, 16 male) aged between four and eight ($M = 6$, $SD = 1.35$) by advertising in local libraries, childcare centers, and through snowball sampling. Participants were randomly allocated to either the experimental condition using our system ($N = 19$) or the control condition involving a human partner ($N = 16$). All the participants’ predominant home language is English. No significant differences were observed between the two conditions in terms of children’s gender, age, baseline math language skills, race/ethnicity, prior use of conversational agents, or caregiver education (Table 3).

4.4 Results

In this section, we discuss our findings regarding children’s learning of math language, their engagement, performance in story creation, as well as enjoyment with either a human partner or AI, MATHEMYTHS. It’s important to note that our participants ranged in age from four to eight years old. This provided a valuable opportunity to examine how children at different developmental stages interact with AI. Thus, we divided the child participants into two age groups: a younger group (ages 4 to 5) and an older group (ages 6 to 8) [18]. This division allowed us to investigate whether and how developmental differences might influence our results.

4.4.1 Learning of Math Language. Our analysis focused on examining if there was an improvement in children’s understanding of targeted math terms from the pretest to the posttest, first considering the effects by condition (i.e., AI versus human) and then examining if these gains varied between the younger and older

age groups. To address these questions, we carried out a two-way repeated-measures mixed ANOVA, using condition and age group as covariates.

As shown in Figure 6, the results indicated significant improvements between the pre- ($M = 16.686$, $SD = 9.578$) and post-test ($M = 18.971$, $SD = 1.526$) math language total scores ($F(1, 31) = 17.009$, $p < .000^{***}$). Moreover, these learning gains were found to be comparable across the AI (pre: $M = 17.105$, $SD = 9.882$; post: $M = 19.500$, $SD = 11.074$) and human (pre: $M = 16.187$, $SD = 9.501$; post: $M = 18.344$, $SD = 10.158$) conditions ($F(1, 31) = 0.012$, $p = .912$). Statistically, children’s learning gains showed marginal variation across different age groups ($F(1, 31) = 3.219$, $p = .083$). On a descriptive level, older children seemed to benefit more from the co-creation activity than younger children, regardless of their interaction with MATHEMYTHS or a human partner. Thus, this age difference did not seem to stem from the nature of interacting with AI, but rather might be due to the fact that older children already have a better understanding of the underlying math concepts in the first place, which might amplify their comprehension of the associated math language.

To further unpack children’s learning gains across each dimension (i.e., definition, recall, transfer, practice), we carried out additional repeated-measures ANOVA analyses. These analyses used experimental conditions and age group as covariates, following the same model we applied in analyzing the total scores. Overall, we found that children’s gain in math language appeared to be primarily driven by their improved scores in answering transfer and practice questions, as the pre-to-post-test gains were statistically significant in these dimensions, but not in definition and recall. However, when analyzing the pre-to-post gains between the story creation partners (human vs AI), AI partners showed a notable advantage in enhancing children’s performance in definition questions ($F(1, 18) = 8.308$, $p = .010^{**}$; pre: $M = 2.737$, $SD = 1.939$; post: $M = 3.368$, $SD = 2.047$), while children with human partners did not show significant improvement in this dimension ($F(1, 15) = 0.015$, $p = .903$; pre: $M = 2.813$, $SD = 1.621$; post: $M = 2.844$, $SD = 1.777$). Further analysis by age group revealed a consistent pattern across all dimensions: older children showed greater improvements than younger ones, although these differences were not statistically significant.

Taking together, MATHEMYTHS has shown to be as effective as a human partner in supporting children’s math language learning through storytelling, while also displaying a stronger advantage in aiding definition comprehension. Moreover, children from both younger and older age groups benefited from this activity, whether interacting with an AI or human, though older children showed a slightly greater advantage.

4.4.2 Engagement. In our study, children actively engaged in the story co-creation activity with MATHEMYTHS. Focusing on the quantity of their engagement (i.e., response rate and length), children responded to over 95% of the questions posed by MATHEMYTHS, with their responses averaging between 6–10 words in length. Generally, children tended to be more responsive and provided longer answers to questions asked by the human partner. Notably, the observed differences in responsiveness and response length between interactions with AI and humans appeared to be predominantly exhibited

Table 3: Participant information by study conditions. An independent t-test and a series of Chi-squared tests revealed that there are no significant differences between the two conditions in terms of baseline math language skills, age, children's gender, race/ethnicity, caregiver education, or prior use of conversational agents.

	Full Sample	AI	Human	Difference
Math Language Baseline	16.686	17.105	16.188	$T(33) = 0.279, p = .391$
Age				$\chi^2(1) = 0.308, p = .579$
Full sample	6.000 ($N = 35$)	6.105 ($N = 19$)	5.875 ($N = 16$)	
4-5-year-olds	4.462 ($N = 13$)	4.333 ($N = 6$)	4.571 ($N = 7$)	
6-8-year-olds	6.909 ($N = 22$)	6.923 ($N = 13$)	6.889 ($N = 9$)	
Gender				$\chi^2(1) = 0.046, p = .830$
Female	54.29% ($N = 19$)	52.63% ($N = 10$)	56.25% ($N = 9$)	
Male	45.71% ($N = 16$)	46.37% ($N = 9$)	43.75% ($N = 7$)	
Race/Ethnicity				$\chi^2(4) = 4.500, p = .343$
Black	11.43% ($N = 4$)	15.79% ($N = 3$)	6.25% ($N = 1$)	
Latino	8.57% ($N = 3$)	5.26% ($N = 1$)	12.50% ($N = 2$)	
White	60.00% ($N = 21$)	68.42% ($N = 13$)	50.00% ($N = 8$)	
Others	17.14% ($N = 6$)	10.53% ($N = 2$)	31.25% ($N = 5$)	
Parents' Education				$\chi^2(2) = 3.240, p = .198$
Above Bachelor's degree	34.29% ($N = 12$)	47.37% ($N = 9$)	20.00% ($N = 3$)	
Bachelor's degree	54.29% ($N = 19$)	42.11% ($N = 8$)	68.75% ($N = 11$)	
Less than Bachelor's degree	11.43% ($N = 4$)	10.53% ($N = 2$)	12.50% ($N = 2$)	
Usage of CA				$\chi^2(4) = 1.689, p = .793$
Daily	17.14% ($N = 6$)	21.05% ($N = 4$)	12.50% ($N = 2$)	
Weekly	17.14% ($N = 6$)	21.05% ($N = 4$)	12.50% ($N = 2$)	
Monthly	14.29% ($N = 5$)	15.79% ($N = 3$)	12.50% ($N = 2$)	
Less than once a month	28.57% ($N = 10$)	21.05% ($N = 4$)	37.50% ($N = 6$)	
Never	22.86% ($N = 8$)	21.05% ($N = 4$)	25.00% ($N = 4$)	
N	35	19	16	

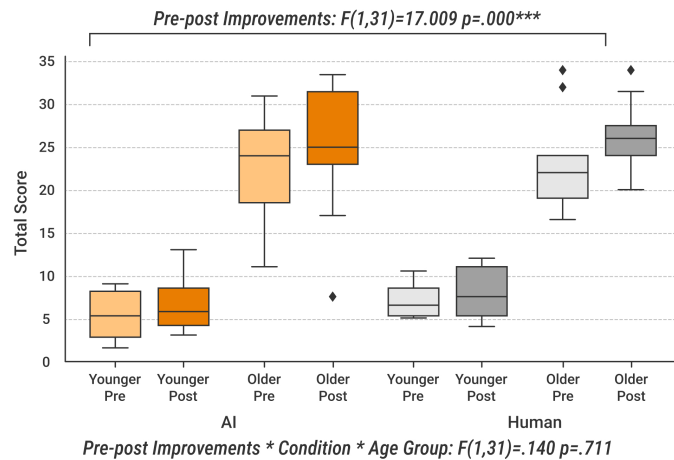


Figure 6: Box plots illustrating the data distribution and the results of a two-way repeated-measures mixed ANOVA for the pre-post-test, using condition and age group as covariates, in the mathematical language assessment. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{}$, $p < 0.001^{***}$.**

by older children. As shown in Figure 7, younger children did not differentiate in their behavior towards the AI versus a human: in both conditions, younger children exhibited almost exactly the same rate (AI: $M = 96.002$, $SD = 5.860$; Human: $M = 96.922$, $SD = 6.511$) and average response length (AI: $M = 8.414$, $SD = 6.579$; Human: $M = 8.968$, $SD = 4.328$). However, older children displayed a different pattern of interaction with humans, characterized by

more frequent (AI: $M = 92.668$, $SD = 13.436$; Human: $M = 99.246$, $SD = 2.456$) and lengthier (AI: $M = 6.064$, $SD = 4.546$; Human: $M = 10.358$, $SD = 7.267$) responses compared to their interactions with the AI. However, this observed pattern did not reach statistical significance: The results of two-way ANOVA analyses indicated that there were no significant differences between younger and older participants across all engagement dimensions.

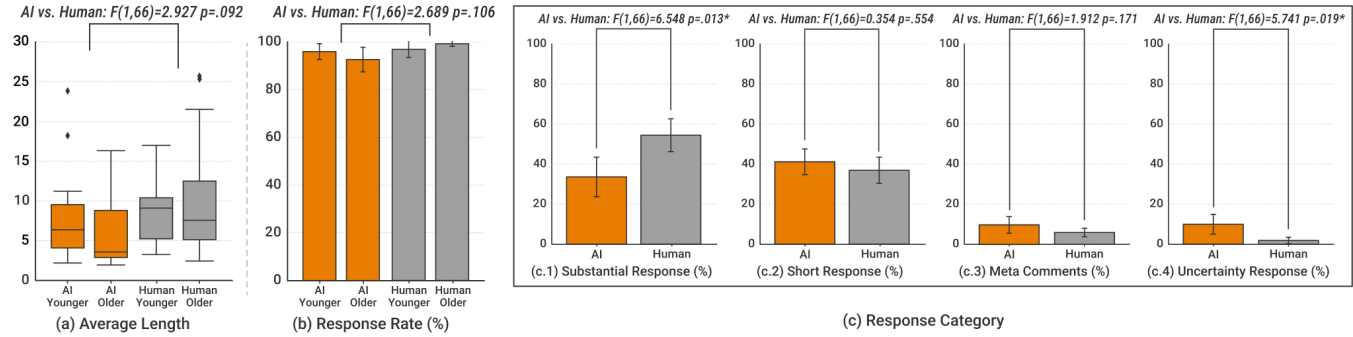


Figure 7: Box plots and bar plots illustrating the distribution of data and the results from the two-way ANOVA regarding children's verbal engagement. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{}$, $p < 0.001^{***}$. Error bars represent 95% confidence intervals (CIs).**

In terms of the nature of children's responses, the majority were actual answers that advanced the stories, either substantial or short. This trend was consistent, regardless of whether children partnered with MATHEMYTHS or a human. However, we observed that children were less likely to provide substantial responses when interacting with MATHEMYTHS compared to a human. Additionally, interactions with MATHEMYTHS elicited a higher incidence of uncertain responses, like "I don't know", though still relatively infrequent, at a rate of 9.9%, as opposed to 1.8% when interacting with humans. This increased frequency of uncertainty in the AI condition might be attributed to MATHEMYTHS's tendency to pose unusual or fantasy-oriented questions. For instance, MATHEMYTHS asked questions like how the characters can use a cloud to speed up their journey, which are less realistic than questions asked by a researcher such as "How can they get through the traffic circle faster?" Nevertheless, the use of unusual and imaginative elements in responses may not necessarily be negative, and is actually a common feature in children's literature, especially for younger audiences. Our analysis of uncertain responses by younger versus older children revealed an interesting trend. While children in general exhibited a higher rate of uncertainty when responding to AI, this tendency was more pronounced among older children. Specifically, younger children showed a 7.6% rate of uncertain responses to AI, compared to 10.9% for older children. In contrast, with human interaction, the rate of uncertain responses was only 3.5% for younger children and even lower at 0.5% for older children.

4.4.3 Story Creation Performance. We then focused on the quality of the stories children created, measured by the number of ideas and elaborations. On average, a child produced about 0.5 idea per prompt during interactions with either AI or humans (AI: $M = .443$, $SD = .392$; Human: $M = .616$, $SD = .404$), amounting to nearly one idea every two prompts. In terms of elaboration, children typically offered 0.3 detailed elaboration per turn, equating to approximately 1 elaborate response every three prompts interactions with both AI and humans (AI: $M = .285$, $SD = .241$; Human: $M = .360$, $SD = .390$). As shown in Figure 8, a two-way ANOVA, examining experimental conditions and age groups, indicated that there were no significant differences in the quality of story creation between the MATHEMYTHS and human conditions (ideas: $F(1, 66) =$

2.380 , $p = .128$; elaboration: $F(1, 66) = .551$, $p = .461$), nor between younger and older children (ideas: $F(1, 66) = .165$, $p = .686$; elaboration: $F(1, 66) = .130$, $p = .720$).

Nevertheless, it might be quite surprising that younger children performed on par with their older peers in this metric, especially for those who interacted with the MATHEMYTHS (Figure 8), despite younger children's presumably less developed language ability. Yet, this might be due to the scaffolding features that we included, such as posing follow-up questions and providing hints when children did not formulate substantive responses. These features appeared to be effective in further eliciting children's responses and engaging them in back-and-forth interactions around one prompt. Thus, the scaffolding features could have reduced the presumed differences in the extent of idea elaboration between older and younger children.

4.4.4 Enjoyment. As shown in Figure 9, children's responses to the survey items indicated an overall positive perception of the story co-creation experience. Descriptively, they showed a more favorable reaction to reading with a human partner in terms of enjoyment (AI: $M = 3.68$, $SD = 1.250$; Human: $M = 4.31$, $SD = 1.195$) and comfort level (AI: $M = 3.42$, $SD = 1.071$; Human: $M = 3.88$, $SD = 1.310$) compared to AI. Moreover, children slightly favored the idea of having the experimenter they interacted with as a reading partner (AI: $M = 3.05$, $SD = 1.682$; Human: $M = 3.38$, $SD = 1.628$) but expressed more interest in reading another story with MATHEMYTHS (AI: $M = 2.42$, $SD = 1.805$; Human: $M = 2.38$, $SD = 1.455$). However, a two-way ANOVA comparing the two conditions across different ages indicated that there were no significant differences between the AI and human conditions ($F(1, 31) = .501$, $p = .484$) or among younger and older participants ($F(1, 31) = .106$, $p = .746$) across all questions in this enjoyment questionnaire. Additionally, no interaction effect was observed between age groups and conditions.

4.4.5 MATHEMYTHS Performance. The findings in terms of children's learning, engagement, story creation performance, and enjoyment above might be better contextualized within the performance of MATHEMYTHS. In this section, we discuss MATHEMYTHS's performance in terms of interpreting children's spoken input as well as its performance in generating responses for children.

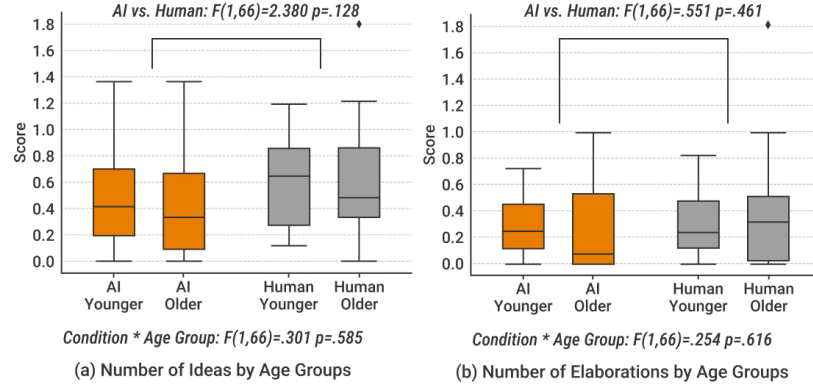


Figure 8: Box plots illustrating the data distribution, along with the results of the two-way ANOVA comparing the story creation performance of children, using condition and age group as covariates. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{}$, $p < 0.001^{***}$.**

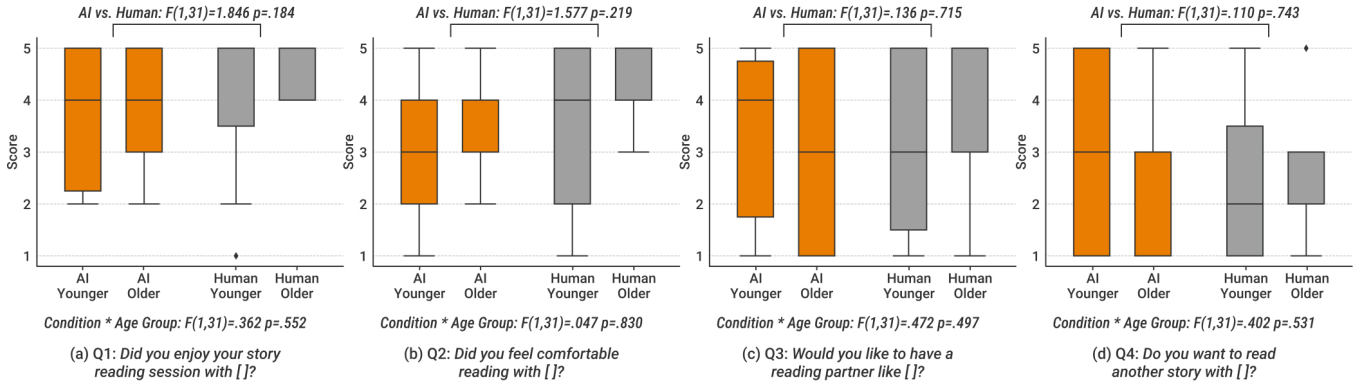


Figure 9: Box plots illustrating the data distribution, along with the results of the two-way ANOVA comparing children's enjoyment, using condition and age group as covariates. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{}$, $p < 0.001^{***}$.**

In terms of speech recognition performance, we manually assessed the accuracy with which the agent in our system recognized children's speech and provided semantically meaningful feedback during each conversational exchange. Our findings revealed that 87% of children's utterances were accurately recognized, resulting in feedback that accurately incorporated children's input. Electing unconventional names presented difficulties for the system to accurately decipher. It is noteworthy that this challenge was not exclusive to our automated system but was also encountered by our human experimenter. Furthermore, certain errors might not be solely ascribed to the limitations of the speech recognition technology; they were related to the lack of sufficient contextual information in the children's speech. For instance, when a child said the word "knight" in isolation, the system registered it as "night" and subsequently proceeded with the follow-up question based on this interpretation. This prompted the child to clarify this with MATHEMYTHS by specifying "A knight with a sword." Among the successful responses, we adapted an open-source, validated gibberish detection model [50] to measure the frequency of nonsensical responses. According to the model's predictions, none of the AI responses were

identified as non-sensible. About 80% were classified as entirely sensible, while the remaining responses were borderline sensible (overall sensible but containing some elements that were illogical).

In terms of the performance of the story creation by AI, we used the same metric as we used to evaluate children's story creation, namely the number of ideas and elaborations. We compared the number of ideas and elaborations observed in AI-generated responses versus responses produced by human researchers. On average, the number of ideas and elaborations generated by humans ($M = 1.938$) and AI ($M = 2.585$) is quite comparable. However, there is a substantially larger variation in the stories generated by humans ($SD = 0.983$) than those generated by AI ($SD = 0.417$), suggesting a higher level of consistency in AI performance. Further examination of the nature of ideas and elaborations generated by the AI and humans corroborated the findings of the model evaluation mentioned in Section 3.3.2. To reiterate, while MATHEMYTHS' ideas and elaborations were comparable to humans in terms of creativity and math relevance, the AI's contributions posed more

challenges in readability, likely attributed to its frequent incorporation of fantasy elements and unrealistic story plots (e.g., a giraffe shrinks itself to fit a small space.)

Taking together, our analysis suggested that while MATHEMYTHS performs satisfactorily in terms of speech recognition and story creation. However, there are certain aspects, such as the characteristics of the narrative stories, diverge from the typical patterns seen with a human story co-creation partner. In the Discussion section below, we will elucidate how such human/AI similarities and differences may have contributed to variations in children's learning approaches and interaction behaviors.

5 DISCUSSION

This paper explored the potential of using LLMs to develop conversational systems for teaching mathematical language via child-AI co-creative storytelling. Our iterative design and development process provides insights into how prompt engineering methods might feasibly be adapted to suit LLMs for young children in a specific educational context, and the model evaluation suggests that prompt-engineered LLMs could produce questions and stories approaching human-like quality. A subsequent user study further showed that the math learning outcomes resulting from interactions with LLMs were also comparable with those from interactions with humans, yet the engagement patterns of different ages differed.

In the remainder of this section, we will discuss how our study speaks to the growing body of literature on child-AI interactions and how we introduce new evidence to still under-explored areas, the design of LLMs as educational, conversational partners. We will then discuss design implications that could guide future developments. Finally, we will address potential limitations and outline a future research agenda.

5.1 AI as Storytelling and Learning Partners

Our paper provides evidence on children's learning and engagement with MATHEMYTHS, as an example of LLMs-based conversational partners. There is a substantial body of prior studies showing that children can learn from their interactions with AI. However, these learning experiences were primarily structured to align with formal pedagogical discourse, in which children were expected to respond to the AI's questions with definitive, correct answers [100, 102, 103]. However, our study goes further to suggest that children can acquire mathematical language through free-form interactions in a narrative-based context with AI, replicating the benefits of the approach used by human educators to teach mathematical language through storytelling. Moreover, our results indicate that this form of learning can manifest in multiple ways. Children demonstrated an improved ability to define and recall the mathematical terms they were exposed to. Moreover, they exhibited proficiency in transferring this knowledge to different contexts and using these terms more accurately. Interestingly, earlier studies that explored the educational benefits of interacting with an AI companion primarily focused on assessing children's information recall [99, 110]. This emphasis may have arisen because those previous studies were intentionally structured to facilitate more organized conversations for a set of facts [99]. However, in our study, we observed that

children not only improved their information recall but also demonstrated enhancements in applying that information and expressing creativity in their language use. This observation can, to some extent, be attributed to the increased sense of empowerment and agency that children experienced during their story co-creation with MATHEMYTHS.

Nonetheless, children appeared to display somewhat different engagement patterns during their interactions with MATHEMYTHS in comparison to their interactions with our human researcher. When considering response quantity, including response rate and length, it appears that children exhibited a slightly more active engagement pattern when interacting with humans. However, this disparity became noticeable primarily among the older children in our study, as younger children did not seem to differentiate in terms of response rate or response length between interactions with AI and humans. This growing divergence in engagement with AI and humans with age could be attributed, in part, to children's perceptions of AI, in particular, their understanding of what AI's capabilities and limitations [31, 104]. Specifically, previous research has identified that older children (aged 7-8) were more inclined to seek factual information from voice assistants and increasingly sought personal information from humans, as compared to younger children (aged 4 and 5) [31]. Intriguingly, the trend observed in our study aligns with age-related findings concerning the behavior of seeking personal information rather than factual information. This alignment is to some extent consistent with the design of MATHEMYTHS, which aims to promoting story co-creation that might deviate from a question-answering pedagogical paradigm where the conversational AI's primary function is to provide factual information. Furthermore, the differences between age groups were also demonstrated in their behavior of using our scaffolding features. Younger children required support from the scaffolding on more occasions, yet their creative performance with the scaffolding was on par with that of their older peers. This finding aligns with expectations, considering younger children possess less-developed literacy skills. Our scaffolding features enable younger children to engage in and benefit from the new math language learning activity on an equal footing with older children. This supports previous research advocating for the initiation of math language interventions at an earlier age [71, 82]. Lastly, we examined children's preferences for co-creating stories with AI or humans and found that children rated both AI and human favorably. However, despite our best efforts to ensure focused evaluations of the specific co-creation partners, it remains essential to acknowledge that children might have drawn upon prior AI interactions, such as those with voice assistants at their homes, which could potentially influence their perceptions during the study [91].

5.2 Design Challenges of LLMs for Educational Technologies

In the design of MATHEMYTHS, we encountered three challenges related to leveraging LLMs for educational technologies.

5.2.1 Iterative Prompt Engineering Optimization for LLMs. The capability of LLMs has attracted much attention from researchers to build educational tools for children [51]. This paper provides pivotal evidence on how to optimize LLMs as story co-creators for

children’s math language learning through prompt engineering. During our development, we found that LLMs, with their impressive generative capabilities, exhibited a higher degree of unpredictability in their outputs compared to specialized machine learning models specifically tailored for narrower tasks. The primary reason is that the nuances of how LLMs interpret our prompts remain a “black-box” for developers. In our efforts at prompt-engineering, there was no guarantee that revising prompts would consistently yield the desired outcomes. This uncertainty aligns with Lin et al.’s research [59], where they reported LLMs exhibit “dispersion” in their potential predictions for a fixed input (i.e., prompt). Empirical evidence suggests that prompt engineers often overgeneralize based on single instances of success or failure when modifying prompts, as noted in a recent paper [106]. Therefore, to improve the chances of creating prompts that yield consistent intended results, extensive testing of prompts is crucial. This involves comparing the consistency of outputs and engaging in a refinement process, often characterized by trial and error. Though the prompt engineering process might potentially involve a large number of iterations, refining prompts is relatively less time-consuming, especially when compared to traditional AI model fine-tuning or training.

5.2.2 Fine-grained Control over LLMs’ Performance. While LLMs’ impressive general language capability allows for easy instruction based on descriptive language, it remains a challenge to provide precise directives to control the output specifically. A significant aspect of this challenge is associated with LLM’s limited capacities in numerical reasoning [107]. For instance, given that MATHEMYTHS was tailored for young children, we aimed to use brief and simple language to facilitate comprehension. However, when we set specific word limits within each sentence for GPT-4 (e.g., “*each sentence should not exceed 10 words*”), the model often disregarded such constraints. In contrast, descriptive instructions without numeric values, like “*keep language brief and child-friendly*,” consistently produced better results. In light of these limitations, other research recommends incorporating mechanisms for quality checks when precision in output is a priority. Feedback loops [54] and validation layers [38] have been suggested to ensure the model’s output aligns with the desired requirements.

Another observation in our development process relates to LLMs occasionally sidestepping structural directives from the original prompt after extended conversations. For example, if GPT-4 is instructed to maintain a specific narrative structure in its responses, it might deviate if these instructions are given only at the beginning. This issue could be attributed to the introduction of non-narrative structural context (e.g., story content, children’s responses) throughout the interaction, which imposes a heavy memory load on LLMs [61]. Shi et al. [85] also found that the inclusion of information irrelevant to problem-solving dramatically decreases LLM performance. To counteract this, we employed injection prompting at the end of each input, a point where LLMs perform best in using longer context [61], strategically placing reminders within ongoing interactions. This approach enhanced response consistency and adherence to the desired structural format in every response.

5.2.3 Mitigation of the Impact of LLM Hallucinations. Hallucinations refer to situations where responses generated by LLMs may initially seem plausible but are actually nonsensical or factually

incorrect [48, 57]. These models, without correct understanding of concepts, appear to be knowledgeable because they present language associated with the concepts, but without logical filtering. In our research, MATHEMYTHS was specifically designed to engage children in story creation involving fantasy elements, rather than tasks involving mathematical counting or calculation. This approach may have, to some extent, mitigated the issues of factual inaccuracies. Nevertheless, MATHEMYTHS still produced content that appeared less feasible in a real-life context. For instance, it occasionally created unusual connections between two common story elements, such as “using clouds to add speed to a journey.” Though the idea of using cloud to speed up might seem imaginative in the context of a fantasy story [109], it might be problematic in contexts such as the teaching of scientific concepts, where unrealistic ideas could lead to confusion or misconceptions. This could also explain why the older children were drawn to the human partner, as they might find the nonsensical language weird. Nevertheless, as the field of hallucination mitigation in LLMs is relatively nascent, future research could benefit from exploring post-processing techniques that scrutinize AI-generated content that might contain hallucinatory information [23] or employing external knowledge as a framework to guide and calibrate the generation of information [78] in LLM-based dialogue systems.

5.3 Design Implications for Child-AI Co-Creative Storytelling Systems

In this section, we discuss some design implications of our study. Our team has already begun improvements to our co-creative storytelling system based on the considerations below.

5.3.1 Generating Adaptive Questions. During storytelling, MATHEMYTHS generates questions to elicit children’s responses. However, some children found certain questions challenging to answer. This was evident from the notably higher rate of uncertainty in their responses to AI-generated questions compared to those posed by human partner. While we implemented scaffolding features to assist children when they struggle to respond, the AI might not always capture the nuances of a child’s emotions and experiences. Specifically, it may not always discern precisely what factors (e.g., language barriers, math knowledge gaps, or unwillingness to participate) caused their struggle. As a result, while the AI provided some scaffolding based on several broad categories, it did not offer more tailored adjustments in response to the unique challenges faced by individual children. One design consideration is to integrate more adaptive algorithms into LLM-based tools that can accurately assess a child’s cognitive load based on their responses and adjust the complexity of the questions accordingly.

5.3.2 Providing Multi-modal Creativity Support. Our study revealed the potential benefits of child-AI co-creative storytelling for enhancing children’s creativity. MATHEMYTHS was developed for verbal interactions with children, which was a significant initial step considering our primary focus on the language capabilities of LLMs. For future designs, incorporating graphical interfaces or multimedia elements could be beneficial. Such additions would enable children to visualize the story narratives, allowing them to express their ideas not only through words but also through drawings, animations,

and sound effects. This can not only augment children’s creative experience [108, 109], but also improve children’s comprehension of abstract concepts [22]. In this context, consistent interactions with AI, which might promote long-term creativity in children, become even more significant as they have shown the capability to learn or emulate creativity from AI interactions [5, 6].

5.3.3 Supporting Embodied Interaction. In this paper, MATHEMYTHS engages children in joint storytelling through a device similar to smart-speakers. While this platform can be easily navigated via voice, it might fall short in conveying emotions or actions—a vital aspect that captivates children during interactions with human counterparts. To compromise it during our user study, we provided children with a plushy. Nonetheless, this might not fully capture the advantages of tangible interactions. A promising solution could be to integrate this virtual agent with an embodied robot, which can offer more vivid feedback through its human-like expressions and behaviors [43, 95]. It is potential that an embodied robot could stimulate different or perhaps heightened levels of engagement. However, the cost implications and scalability concerns associated with robots cannot be ignored.

5.4 Limitations and Future Work

While our study offers preliminary insights, it also paves the way for addressing further questions in the future.

First, in our user study, we compared children’s interactions with MATHEMYTHS with those with a trained researcher who adhered to a strict protocol for narrating stories. This protocol ensured the inclusion of a specified number of mathematical terms. While this approach granted us a controlled environment, ensuring consistent mathematical language exposure for the children, it might have constrained the natural adaptability and spontaneity that a human experimenter can bring to the storytelling process. In future research, it would be intriguing to also compare children’s engagement and enjoyment with a human partner who can fully harness their natural storytelling capabilities.

Related, we compared the learning outcomes between storytelling with AI or humans but did not include a comparison to more conventional instructional methods, such as directly teaching the terms, providing examples, allowing the child to practice, and offering corrective feedback. Although prior research has indicated superior learning outcomes with story-based methods involving a teacher, exploring whether this advantage over direct instruction extends to AI-driven contexts presents an intriguing research avenue. Additionally, our current participant pool predominantly consists of families with higher educational backgrounds. These children are likely to have greater access to conversational AI technologies, which may have influenced their interaction with our system. Future research should focus on including children with limited technological access to evaluate if LLM-based learning tools remain effective in such contexts.

In addition, our study’s participants engaged in only two AI collaborative storytelling sessions. This short-term access might not fully reflect the dynamics of extended interactions children might have with such systems. As children become more accustomed to the AI over time, their expectations could evolve, potentially requiring more advanced and adaptive storytelling experiences. On

the other hand, repeated sessions could lead to a more harmonized child-AI collaboration, potentially enriching the storytelling process. Future studies should explore these longer-term dynamics, perhaps by giving children extended opportunities to interact with the AI storytelling system and tracking their engagement patterns over extended periods.

5.5 A Note on Ethical Considerations

The safe use of AI by children has been a topic of public discussion for some time. As we transition to an era dominated by generative AI, the debate becomes even more pressing. In this new landscape, children may directly encounter content produced by AI without the protective filters traditionally provided by human oversight. In this section, we will focus on some ethical considerations pertaining to privacy and safety.

In terms of privacy, MATHEMYTHS utilizes two cloud-based services to process children’s utterances: Google’s speech-to-text for transcribing and OpenAI’s GPT-4 to interpret their responses. Both companies offer measures to protect children’s data. Specifically, children’s utterances are not stored and are promptly deleted after transcription by Google Cloud. Additionally, OpenAI commits not to train their models with data passed through their APIs⁷. However, when these cloud-based AI products enter the market, the adequacy of children’s data privacy protection remains uncertain due to the absence of specific regulations for AI products. While the US Children’s Online Privacy Protection Act (COPPA) regulates the collection and use of personal information from those under 13, it has yet to be updated to tackle the emerging concerns related to AI.

In terms of safety, to safeguard against inappropriate AI-generated content, we integrated a moderation layer, utilizing the OpenAI APIs⁸, to ensure outputs were devoid of sexual content, hate speech, harassment, violence, or self-harm. However, this content filtering approach might not be as effective in addressing biases and stereotypes (e.g., when a child asks “What do girls do?”, and the AI agent responds “housewives”), given that these biases are as deeply ingrained in the data used to train AI systems as they are in the rest of the culture. Nonetheless, a growing number of studies have begun to investigate methods for detecting and reducing such stereotypes and biases, although these studies have not specifically focused on children. Given the present state of technology, a viable approach would be to deploy LLMs within a “walled garden”—limiting interactions to narrower domains, like the math-focused narratives in our study. Such confines might mitigate risks while still capitalizing on the capabilities of LLMs.

6 CONCLUSION

While AI has been utilized for structured interactions with children for quite a while, such as dialogic question-answering, the affordance provided by LLMs in facilitating free-form, creative tasks, like joint storytelling, signifies a groundbreaking shift. This research explores that very frontier. Our findings suggested that, although the learning outcomes with AI mirror those with human partners, the interaction dynamics are different. Instead of viewing this as a

⁷<https://openai.com/api-data-privacy>

⁸<https://platform.openai.com/docs/api-reference/moderations>

limitation, it might be more fruitful to recognize AI as providing a unique interaction opportunity. These interactions, while different, could supplement children's everyday human interactions. Given the rapid advancements in LLMs, the capabilities we observe today are likely to further refine in the near future. Yet, the insights garnered now will shape and anchor our design philosophies as we move forward.

ACKNOWLEDGMENTS

This paper is supported by the National Science Foundation under Grant No. 2302730. We thank the children who participated in our study and our local public libraries for graciously offering space for us to recruit participants and carry out the study. We also thank the research assistants at the University of Michigan for their assistance with data collection and analysis.

REFERENCES

- [1] Rania Abdelghani, Yen-Hsiang Wang, Xingdi Yuan, Tong Wang, Pauline Lucas, Hélène Sauzéon, and Pierre-Yves Oudeyer. 2023. GPT-3-Driven Pedagogical Agents for Training Children's Curious Question-Asking Skills. <https://doi.org/10.1007/s40593-023-00340-7> arXiv:2211.14228 [cs]
- [2] Olaiya Aina. 1999. The Importance of Oral Storytelling in Literacy Development. *Ohio Reading Teacher* 33, 1 (1999), 15–18.
- [3] Murat Akkus. 2016. The Common Core State Standards for Mathematics. *International Journal of Research in Education and Science* 2, 1 (2016), 49–54.
- [4] Amal Alabdulkarim, Siyan Li, and Xiangyu Peng. 2021. Automatic Story Generation: Challenges and Attempts. <https://doi.org/10.48550/arXiv.2102.12634> arXiv:2102.12634 [cs]
- [5] Safinah Ali, Tyler Moroso, and Cynthia Breazeal. 2019. Can Children Learn Creativity from a Social Robot?. In *Proceedings of the 2019 on Creativity and Cognition (C&C '19)*. Association for Computing Machinery, New York, NY, USA, 359–368. <https://doi.org/10.1145/3325480.3325499>
- [6] Safinah Ali, Hae Won Park, and Cynthia Breazeal. 2020. Can Children Emulate a Robotic Non-Player Character's Figural Creativity?. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY '20)*. Association for Computing Machinery, New York, NY, USA, 499–509. <https://doi.org/10.1145/3410404.3414251>
- [7] Minhui Bao. 2019. Can Home Use of Speech-Enabled Artificial Intelligence Mitigate Foreign Language Anxiety—Investigation of a Concept. *Arab World English Journal (AWEJ) Special Issue on CALL* 1, 5 (2019), 28–40.
- [8] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2020. Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.
- [9] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.
- [10] Marina Umaschi Bers and Justine Cassell. 1998. Interactive Storytelling Systems for Children: Using Technology to Explore Language and Identity. *Journal of Interactive Learning Research* 9 (1998), 183–215.
- [11] Marina Umaschi Bers and Justine Cassell. 1998. Interactive storytelling systems for children: Using technology to explore language and identity. *Journal of Interactive Learning Research* 9 (1998), 183–215.
- [12] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. <https://doi.org/10.48550/arXiv.2005.14165> arXiv:2005.14165 [cs]
- [13] Adriana G. Bus, Marinus H. van IJzendoorn, and Anthony D. Pellegrini. 1995. Joint Book Reading Makes for Success in Learning to Read: A Meta-Analysis on Intergenerational Transmission of Literacy. *Review of Educational Research* 65, 1 (March 1995), 1–21. <https://doi.org/10.3102/00346543065001001>
- [14] John Bynner and Samantha Parsons. 1997. *Does Numeracy Matter? Evidence from the National Child Development Study on the Impact of Poor Numeracy on Adult Life*. Technical Report. Basic Skills Agency, Commonwealth House, 1-19 New Oxford Street, London WC1A 1NU, England, United Kingdom (6).
- [15] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ip-polito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivan Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. <https://doi.org/10.48550/arXiv.2204.02311> arXiv:2204.02311 [cs]
- [16] John Joon Young Chung, Woosok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–19. <https://doi.org/10.1145/3491102.3501819>
- [17] Eric Dearing, Beth Casey, Pamela E Davis-Kean, Sarah Eason, Elizabeth Gundersen, Susan C Levine, Elida V Laski, Melissa Libertus, Linxi Lu, Caitlin McPherran Lombardi, et al. 2022. Socioeconomic variations in the frequency of parent number talk: A meta-analysis. *Education Sciences* 12, 5 (2022), 312.
- [18] Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L Murnane, and James A. Landay. 2021. StoryCoder: Teaching Computational Thinking Concepts Through Storytelling in a Voice-Guided App for Children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3411764.3445039>
- [19] R. Dijkstra, Z. Geng, S. Kayal, and J. Kamps. 2022. Reading Comprehension Quiz Generation Using Generative Pre-Trained Transformers. In *The 23th International Conference on Artificial Intelligence in Education (AIED)*. Springer, New York, NY, USA, 1–14.
- [20] Greg J. Duncan, Chantelle J. Dowsett, Amy Claessens, Katherine Magnuson, Aletha C. Huston, Pamela Klebanov, Linda S. Pagani, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, Holly Sexton, Kathryn Duckworth, and Crista Japel. 2007. School Readiness and Later Achievement. *Developmental Psychology* 43, 6 (Nov. 2007), 1428–1446. <https://doi.org/10.1037/0012-1649.43.6.1428>
- [21] Alessandro Duranti and Charles Goodwin. 1992. *Rethinking Context: Language as an Interactive Phenomenon*. Number 11 in 1. Cambridge University Press, Cambridge, England.
- [22] Utkarsh Dwivedi, Nitendra Rajput, Prasenjit Dey, and Blessin Varkey. 2017. VisualMath: An Automated Visualization System for Understanding Math Word Problems. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion (IUI '17 Companion)*. Association for Computing Machinery, New York, NY, USA, 105–108. <https://doi.org/10.1145/3030024.3040989>
- [23] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021. Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2197–2214. <https://doi.org/10.18653/v1/2021.emnlp-main.168>
- [24] Reham El Shazly. 2021. Effects of Artificial Intelligence on English Speaking Anxiety and Speaking Performance: A Case Study. *Expert Systems* 38, 3 (2021), e12667.
- [25] Leanne Elliott and Heather J Bachman. 2018. SES disparities in early math abilities: The contributions of parents' math cognitions, practices to support math, and math talk. *Developmental Review* 49 (2018), 1–15.
- [26] Leanne Elliott, Emily J Braham, and Melissa E Libertus. 2017. Understanding sources of individual variability in parents' number talk with young children. *Journal of Experimental Child Psychology* 159 (2017), 1–15.
- [27] Amaechi Uneke Enyi. 2015. Language and Interactional Discourse: Deconstructing the Talk-Generating Machinery in Natural Conversation. *Advances in Language and Literary Studies* 6, 4 (2015), 171–178.
- [28] Erin Elizabeth Flynn. 2018. Ideas in dialogue: Leveraging the power of child-led storytelling in the multicultural preschool classroom. *Language in Society* 47, 4 (2018), 601–633.
- [29] Ebrahim Gabajiwala, Priyav Mehta, Ritik Singh, and Reeta Koshy. 2022. Quiz Maker: Automatic Quiz Generation from Text Using NLP. In *Futuristic Trends in Networks and Computing Technologies: Select Proceedings of Fourth International Conference on FTNCT 2021*. Springer, New York, NY, USA, 523–533.
- [30] Asma Ghandeharioun, Judy Hanwen Shen, Natasha Jaques, Craig Ferguson, Noah Jones, Agata Lapedriza, and Rosalind Picard. 2019. Approximating Interactive Human Evaluation with Self-Play for Open-Domain Dialog Systems. <https://doi.org/10.48550/arXiv.1906.09308> arXiv:1906.09308 [cs, stat]

- [31] Lauren N Girouard-Hallam and Judith H Danovitch. 2022. Children's trust in and learning from voice assistants. *Developmental Psychology* 58, 4 (2022), 646.
- [32] Manuela Glaser, Bärbel Garsoffky, and Stephan Schwan. 2009. Narrative-Based Learning: Possible Benefits and Problems. *The European Journal of Communication Research* 34, 4 (Dec. 2009), 429–447. <https://doi.org/10.1515/COMM.2009.026>
- [33] Arthur C. Graesser, Mark W. Conley, and Andrew Olney. 2012. Intelligent Tutoring Systems. *APA educational psychology handbook* 3 (2012), 451–473.
- [34] Arthur C. Graesser, Sidney D'Mello, Xiangen Hu, Zhiqiang Cai, Andrew Olney, and Brent Morgan. 2012. AutoTutor. In *Applied Natural Language Processing: Identification, Investigation and Resolution*. IGI Global, Hershey, PA, USA, 169–187.
- [35] Jian Guan, Fei Huang, Zhihao Zhao, Xiaoyan Zhu, and Minlie Huang. 2020. A Knowledge-Enhanced Pretraining Model for Commonsense Story Generation. *Transactions of the Association for Computational Linguistics* 8 (Jan. 2020), 93–108. https://doi.org/10.1162/tacl_a_00302
- [36] Brenna Hassinger-Das, Nancy C. Jordan, and Nancy Dyson. 2015. Reading Stories to Learn Math. *The Elementary School Journal* 116, 2 (Dec. 2015), 242–264. <https://doi.org/10.1086/683986>
- [37] Brenna Hassinger-Das, Nancy C. Jordan, and Nancy Dyson. 2015. Reading Stories to Learn Math: Mathematics Vocabulary Instruction for Children with Early Numeracy Difficulties. *The Elementary School Journal* 116, 2 (2015), 242–264.
- [38] Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. 2023. Solving Math Word Problems by Combining Language Models With Symbolic Solvers. <https://doi.org/10.48550/arXiv.2304.09102> arXiv:2304.09102 [cs]
- [39] Alexis Hiniker, Hyewon Suh, Sabina Cao, and Julie A. Kientz. 2016. Screen Time Tantrums: How Families Manage Screen Media Experiences for Toddlers and Preschoolers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 648–660. <https://doi.org/10.1145/2858036.2858278>
- [40] Hui-Ru Ho, Nathan Thomas White, Edward M. Hubbard, and Bilge Mutlu. 2023. Designing Parent-Child-Robot Interactions to Facilitate In-Home Parental Math Talk with Young Children. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference (IDC '23)*. Association for Computing Machinery, New York, NY, USA, 355–366. <https://doi.org/10.1145/3585088.3589358>
- [41] Robin L. Hojnosi, Helen Lynn Columba, and Joy Polignano. 2014. Embedding Mathematical Dialogue in Parent-Child Shared Book Reading: A Preliminary Investigation. *Early Education and Development* 25, 4 (May 2014), 469–492. <https://doi.org/10.1080/10409289.2013.810481>
- [42] Layne Jackson Hubbard, Yifan Chen, Eliana Colunga, Pilyoung Kim, and Tom Yeh. 2021. Child-Robot Interaction to Integrate Reflective Storytelling Into Creative Play. In *Proceedings of the 13th Conference on Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3450741.3465254>
- [43] Layne Jackson Hubbard, Yifan Chen, Eliana Colunga, Pilyoung Kim, and Tom Yeh. 2021. Child-Robot Interaction to Integrate Reflective Storytelling Into Creative Play. In *Creativity and Cognition (C&C '21)*. Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3450741.3465254>
- [44] Charlotte S. Huck and Barbara Zulantz Kiefer. 2003. *Children's Literature in the Elementary School* (8th edition ed.). McGraw-Hill College, Boston.
- [45] Common Core State Standards Initiative. 2010. Mathematics Standards – Common Core State Standards Initiative.
- [46] Clara M. Jennings, James E. Jennings, Joyce Richey, and Lisbeth Dixon-Krauss. 1992. Increasing Interest and Achievement in Mathematics through Children's Literature. *Early Childhood Research Quarterly* 7, 2 (June 1992), 263–276. [https://doi.org/10.1016/0885-2006\(92\)90008-M](https://doi.org/10.1016/0885-2006(92)90008-M)
- [47] Hyangeun Ji, Insook Han, and Yujung Ko. 2022. A Systematic Review of Conversational AI in Language Education: Focusing on the Collaboration with Human Teachers. *Journal of Research on Technology in Education* 55, 1 (2022), 48–63.
- [48] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (March 2023), 248:1–248:38. <https://doi.org/10.1145/3571730>
- [49] Qinyin Jia, Jialin Cui, Yunkai Xiao, Chengyuan Liu, Parvez Rashid, and Edward F. Gehringer. 2021. ALL-IN-ONE: Multi-Task Learning BERT Models for Evaluating Peer Assessments. <https://doi.org/10.48550/arXiv.2110.03895> arXiv:2110.03895 [cs]
- [50] Madhur Jindal. 2022. Madhurjindal/Autonlp-Gibberish-Detector-492513457 · Hugging Face. <https://huggingface.co/madhurjindal/autonlp-Gibberish-Detector-492513457>
- [51] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education. <https://doi.org/10.35542/osf.io/5er8f>
- [52] Saba Kawas, Nicole S. Kuhn, Kyle Sorsokke, Emily Bascom, Alexis Hiniker, and Katie Davis. 2021. When Screen Time Isn't Screen Time: Tensions and Needs Between Tweens and Their Parents During Nature-Based Exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3411764.3445142>
- [53] Anne Lamott. 1995. *Bird by Bird: Some Instructions on Writing and Life* (1st edition ed.). Anchor, New York.
- [54] Dongyub Lee, Taesun Whang, Chanhee Lee, and Heuseok Lim. 2023. Towards Reliable and Fluent Large Language Models: Incorporating Feedback Learning Loops in QA Systems. <https://doi.org/10.48550/arXiv.2309.06384> arXiv:2309.06384 [cs]
- [55] Yoonjoo Lee, Tae Soo Kim, Minsuk Chang, and Juho Kim. 2022. Interactive Children's Story Rewriting Through Parent-Children Interaction. In *Proceedings of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022)*. Association for Computational Linguistics, Dublin, Ireland, 62–71. <https://doi.org/10.18653/v1/2022.in2writing-1.9>
- [56] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. <https://doi.org/10.1145/3544548.3581369>
- [57] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaliEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models. <https://doi.org/10.48550/arXiv.2305.11747> arXiv:2305.11747 [cs]
- [58] Margaret Li, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-Turn Comparisons. <https://doi.org/10.48550/arXiv.1909.03087> arXiv:1909.03087 [cs]
- [59] Zhen Lin, Shubendu Trivedi, and Jimeng Sun. 2023. Generating with Confidence: Uncertainty Quantification for Black-Box Large Language Models. <https://doi.org/10.48550/arXiv.2305.19187> arXiv:2305.19187 [cs, stat]
- [60] Ellen C. Litkowski, Robert J. Duncan, Jessica A.R. Logan, and David J. Purpura. 2020. When do preschoolers learn specific mathematics skills? Mapping the development of early numeracy knowledge. *Journal of Experimental Child Psychology* 195 (2020), 104846.
- [61] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. <https://doi.org/10.48550/arXiv.2307.03172> arXiv:2307.03172 [cs]
- [62] Silvia B. Lovato and Anne Marie Piper. 2019. Young Children and Voice Search: What We Know from Human-Computer Interaction Research. *Frontiers in psychology* 10 (2019), 8.
- [63] Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. Association for Computing Machinery, New York, NY, USA, 301–313.
- [64] Kurt Messick. 2023. *ChatGPT as Author - 100 Words or Less: 100 Short Stories Generated by AI*. Independently published, Bloomington, Indiana, USA.
- [65] Lesley Mandel Morrow. 1985. Retelling Stories: A Strategy for Improving Young Children's Comprehension, Concept of Story Structure, and Oral Language Complexity. *The Elementary School Journal* 85, 5 (May 1985), 647–661. <https://doi.org/10.1086/461427>
- [66] Olga Nelson. 1989. Storytelling: Language experience for meaning making. *The Reading Teacher* 42, 6 (1989), 386–390.
- [67] OpenAI. 2022. Introducing ChatGPT. <https://openai.com/blog/chatgpt>.
- [68] OpenAI. 2023. GPT-4. <https://openai.com/gpt-4>.
- [69] Hiroyuki Osono, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: AI Supported Story Co-Creation via Unsupervised Multitask Learning to Increase Writers' Creativity in Japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–10. <https://doi.org/10.1145/3411763.3450391>
- [70] Alison H. Paris and Scott G. Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly* 38, 1 (2003), 36–76.
- [71] Samantha Parsons and John M. Bynner. 2005. *Does Numeracy Matter More?* National Research and Development Centre for Adult Literacy and Numeracy, London, UK.
- [72] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3174214>
- [73] Sarah R. Powell and Melissa K. Driver. 2015. The Influence of Mathematics Vocabulary Instruction Embedded Within Addition Tutoring for First-Grade Students With Mathematics Difficulty. *Learning Disabilities Quarterly* 38, 4 (Nov. 2015), 221–233. <https://doi.org/10.1177/0731948714564574>
- [74] David J. Purpura and Jessica A. R. Logan. 2015. The Nonlinear Relations of the Approximate Number System and Mathematical Language to Early Mathematics Development. *Developmental Psychology* 51, 12 (2015), 1717–1724. <https://doi.org/10.1037/dev0000055>

- [75] David J. Purpura, Amy R. Napoli, and Yemimah King. 2019. Development of Mathematical Language in Preschool and Its Role in Learning Numeracy Skills. In *Cognitive Foundations for Improving Mathematical Learning*. Elsevier Academic Press, San Diego, CA, US, 175–193. <https://doi.org/10.1016/B978-0-12-815952-1.00007-4>
- [76] David J. Purpura, Amy R. Napoli, Elizabeth A. Wehrspann, and Zachary S. Gold. 2017. Causal Connections Between Mathematical Language and Mathematical Knowledge: A Dialogic Reading Intervention. *Journal of Research on Educational Effectiveness* 10, 1 (Jan. 2017), 116–137. <https://doi.org/10.1080/19345747.2016.1204639>
- [77] David J. Purpura, Sara A. Schmitt, Amy R. Napoli, Jennifer Dobbs-Oates, Yemimah A. King, Caroline Byrd Hornburg, Lauren Westerberg, Giulia A. Borriello, Lindsey M. Bryant, Laura Y. Anaya, Melody Kung, Ellen Litkowski, Joyce Lin, and Emily Rolan. 2021. Engaging Caregivers and Children in Picture Books: A Family-Implemented Mathematical Language Intervention. *Journal of Educational Psychology* 113 (2021), 1338–1353. <https://doi.org/10.1037/edu0000662>
- [78] Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the Factual Knowledge Boundary of Large Language Models with Retrieval Augmentation. <https://doi.org/10.48550/arXiv.2307.11019> arXiv:2307.11019 [cs]
- [79] Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianqiao Xu, Abdallah AbuHashem, Griffin Dietz, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2020. Supporting Children's Math Learning with Feedback-Augmented Narrative Technology. In *Proceedings of the Interaction Design and Children Conference (IDC '20)*. Association for Computing Machinery, New York, NY, USA, 567–580. <https://doi.org/10.1145/3392063.3394400>
- [80] Kimiko Ryokai, Cati Vaucelle, and Justine Cassell. 2003. Virtual Peers as Partners in Storytelling and Literacy Learning. *Journal of computer assisted learning* 19, 2 (2003), 195–208.
- [81] Michael Sailer, Elisabeth Bauer, Riikka Hofmann, Jan Kiesewetter, Julia Glas, Iryna Gurevych, and Frank Fischer. 2023. Adaptive Feedback from Artificial Neural Networks Facilitates Pre-Service Teachers' Diagnostic Reasoning in Simulation-Based Learning. *Learning and Instruction* 83 (2023), 101620.
- [82] Julie Sarama and Douglas H. Clements. 2009. *Early Childhood Mathematics Education Research: Learning Trajectories for Young Children*. Routledge, New York, NY, USA.
- [83] Michael Schiro. 2004. *Oral Storytelling and Teaching Mathematics: Pedagogical and Multicultural Perspectives*. SAGE, Thousand Oaks, CA, USA.
- [84] Hanieh Shakeri, Carman Neustaedt, and Steve DiPaola. 2021. SAGA: Collaborative Storytelling with GPT-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '21)*. Association for Computing Machinery, New York, NY, USA, 163–166. <https://doi.org/10.1145/3462204.3481771>
- [85] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. <https://doi.org/10.48550/arXiv.2302.00093> arXiv:2302.00093 [cs]
- [86] Prentice Starkey, Alice Klein, and Ann Wakeley. 2004. Enhancing Young Children's Mathematical Knowledge through a Pre-Kindergarten Mathematics Intervention. *Early Childhood Research Quarterly* 19, 1 (2004), 99–120.
- [87] Ming Sun, Iolanda Leite, Jill Fain Lehman, and Boyang Li. 2017. Collaborative Storytelling between Robot and Child: A Feasibility Study. In *Proceedings of the 2017 Conference on Interaction Design and Children (IDC '17)*. Association for Computing Machinery, New York, NY, USA, 205–214. <https://doi.org/10.1145/3078072.3079714>
- [88] Tzu-Yu Tai and Howard Hao-Jan Chen. 2020. The Impact of Google Assistant on Adolescent EFL Learners' Willingness to Communicate. *Interactive Learning Environments* 31, 3 (2020), 1–18.
- [89] Jyothi Thippana, Leanne Elliott, Sarah Gehman, Klaus Libertus, and Melissa E. Libertus. 2020. Parents' use of number talk with young children: Comparing methods, family factors, activity contexts, and relations to math skills. *Early Childhood Research Quarterly* 53 (2020), 249–259.
- [90] Eylül Turan and Bert De Smedt. 2022. Mathematical Language and Mathematical Abilities in Preschool: A Systematic Literature Review. *Educational Research Review* 36 (June 2022), 100457. <https://doi.org/10.1016/j.edurev.2022.100457>
- [91] Jessica Van Brummelen, Viktoriya Tabunshchik, and Tommy Heng. 2021. "Alexa, Can I Program You?": Student Perceptions of Conversational Artificial Intelligence Before and After Programming Alexa. In *Proceedings of the 20th Annual ACM Interaction Design and Children Conference (IDC '21)*. Association for Computing Machinery, New York, NY, USA, 305–313. <https://doi.org/10.1145/3459990.3460730>
- [92] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle. *Journal of Experimental Social Psychology* 52 (May 2014), 113–117. <https://doi.org/10.1016/j.jesp.2014.01.005>
- [93] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. <https://doi.org/10.48550/arXiv.2206.07682> arXiv:2206.07682 [cs]
- [94] Jacqueline Kory Westlund and Cynthia Breazeal. 2015. The Interplay of Robot Language Level with Children's Language Learning during Storytelling. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. IEEE, New York, NY, USA, 65–66.
- [95] Nathan Thomas White, Bengisu Cagiltay, Joseph E Michaelis, and Bilge Mutlu. 2021. Designing Emotionally Expressive Social Commentary to Facilitate Child-Robot Interaction. In *Interaction Design and Children (IDC '21)*. Association for Computing Machinery, New York, NY, USA, 314–325. <https://doi.org/10.1145/3459990.3460714>
- [96] Miranda Kit-Yi Wong and Wing Chee So. 2016. Spoken Narrative Assessment: A Supplementary Measure of Children's Creativity. *Creativity Research Journal* 28, 4 (Oct. 2016), 471–477. <https://doi.org/10.1080/10400419.2016.1229989>
- [97] Chun-Cheng Wu, Chih-Wei Chang, Baw-Jhiune Liu, and Gwo-Dong Chen. 2008. Improving Vocabulary Acquisition by Designing a Storytelling Robot. In *2008 Eighth IEEE International Conference on Advanced Learning Technologies*. IEEE, New York, NY, USA, 498–500.
- [98] Wenjie Xu, Jiayi Ma, Jiayu Yao, Weijia Lin, Chao Zhang, Xuanhe Xia, Nan Zhuang, Shitong Weng, Xiaoqian Xie, Shuyue Feng, Fangtian Ying, Preben Hansen, and Cheng Yao. 2023. MathKingdom: Teaching Children Mathematical Language Through Speaking at Home via a Voice-Guided Game. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3544548.3581043>
- [99] Ying Xu. 2023. Talking with machines: Can conversational technologies serve as children's social partners? *Child Development Perspectives* 17, 1 (2023), 53–58.
- [100] Ying Xu, Joseph Aubele, Valery Vigil, Andres S. Bustamante, Young-Suk Kim, and Mark Warschauer. 2022. Dialogue with a Conversational Agent Promotes Children's Story Comprehension via Enhancing Engagement. *Child Development* 93, 2 (2022), e149–e167. <https://doi.org/10.1111/cdev.13708>
- [101] Ying Xu, Stacy Branham, Xinwei Deng, Penelope Collins, and Mark Warschauer. 2021. Are Current Voice Interfaces Designed to Support Children's Language Development?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–12. <https://doi.org/10.1145/3411764.3445271>
- [102] Ying Xu, Valery Vigil, Andres S. Bustamante, and Mark Warschauer. 2022. "Elinor's Talking to Me!": Integrating Conversational AI into Children's Narrative Science Programming. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. <https://doi.org/10.1145/3491102.3502050>
- [103] Ying Xu and Mark Warschauer. 2020. Exploring Young Children's Engagement in Joint Reading with a Conversational Agent. In *Proceedings of the Interaction Design and Children Conference (IDC '20)*. Association for Computing Machinery, New York, NY, USA, 216–228. <https://doi.org/10.1145/3392063.3394417>
- [104] Ying Xu and Mark Warschauer. 2020. What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376416>
- [105] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. Association for Computing Machinery, New York, NY, USA, 841–852. <https://doi.org/10.1145/3490099.3511105>
- [106] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3544548.3581388>
- [107] Bowen Zhang and Harold Soh. 2023. Large Language Models as Zero-Shot Human Models for Human-Robot Interaction. <https://doi.org/10.48550/arXiv.2303.03548> arXiv:2303.03548 [cs]
- [108] Chao Zhang, Cheng Yao, Jianhui Liu, Zili Zhou, Weilin Zhang, Lijuan Liu, Fangtian Ying, Yijun Zhao, and Guanyun Wang. 2021. StoryDrawer: A Co-Creative Agent Supporting Children's Storytelling through Collaborative Drawing. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, 1–6.
- [109] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child-AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3491102.3501914>
- [110] Chao Zhang, Zili Zhou, Yajing Hu, Lanjing Liu, Jiayi Wu, Yaping Shao, Jianhui Liu, Lingyan Zhang, Lijuan Liu, Hangyue Chen, Fangtian Ying, and Cheng Yao. 2023. Observe It, Draw It: Scaffolding Children's Observations of Plant

Biodiversity with an Interactive Drawing Tool. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference (IDC '23)*. Association for Computing Machinery, New York, NY, USA, 253–266. <https://doi.org/10.1145/3585088.3589380>

- [111] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. StoryBuddy: A Human-AI Collaborative Chatbot for Parent-Child Interactive Storytelling with Flexible Parental Involvement. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–21. <https://doi.org/10.1145/3491102.3517479>

A EXAMPLE CONVERSATION SNIPPETS BETWEEN A CHILD AND MATHEMYTHS

A.1 Warming Up

AGENT: Hello, my little friend! I'm Mathemyths. I'm really looking forward to spending some time with you today! May I know your name, please?

CHILD: My name is Nick.

AGENT: Nice to meet you, Nick! Now, could you please tell me how you're feeling today? Are you feeling happy, excited, or maybe a little sleepy?

CHILD: I'm happy!

AGENT: Thank you for sharing your feelings! I'm here to make your day more fun. How about we make up a story together? We'll use our imaginations and even learn some math words. I'll start, then you can add to the story. If you have any questions or need help, just ask me. Always remember, the power of our imaginations knows no bounds!

A.2 Global Setting

AGENT: Now, let's come up with a cool main character for the story. This could be an animal, a person, a mythical creature, or anything else that fits our story. What do you want our main character to be?

CHILD: The main character is a panda!

AGENT: That's interesting! What's the character's name?

CHILD: Her name is Lucy.

AGENT: Great! Where will our story happen?

CHILD: It will happen in a forest.

AGENT: Alright! We'll dive into an incredible adventure with a panda named Lucy in the forest. Ready to start the story? Let's go!

A.3 Opening

AGENT: Once upon a time, in the heart of a green forest, lived a playful panda named Lucy. One sunny afternoon, she found an unusual map, and along with her new friend, a little child named Nick, they embarked on an exciting journey to find a hidden treasure.

A.4 Storytelling

AGENT: Lucy and Nick decided to ask help from Lucy's father, Mr. Bamboo. He looked at the map and told them, "The treasure is hidden 100 steps away from the ancient oak tree." He then added, "You must also find the half-way point, which is **half** of the total steps, that is, 50 steps from the start."

AGENT: In this context, "**half**" is when you split something into two equal parts, like splitting 100 steps into 50 steps.

AGENT: How will Lucy and Nick to find the **half**-way point?

CHILD: Lucy goes to the halfway point; I go to the end, and finally we find all the treasure!

AGENT: Wonderful idea! That's a clever way to divide and conquer the search.

A.5 Scaffolding

AGENT: What will Lucy and Nick do next to make sure their steps equals the total distance to the treasure?

CHILD: I don't know.

AGENT: That's okay! Let's think together. Lucy and Nick can take turns counting their steps. If they keep track, how can they make sure their total steps equal the distance to the treasure?

B PROCESS OF CRAFTING DATASET FOR EVALUATING STORY CONTINUATIONS

To evaluate whether our prompt-engineered GPT-4 model could continue the story adaptively based on children's contributions and integrate mathematical language within the story plot, we compared the GPT-4 generated stories with those generated by humans as our baseline. GPT-4 generated stories were created using a self-chat technique [30], which enabled our dialogue system to engage in self-generated conversations, mimicking a child's interactions, and collaboratively creating stories. From this, we compiled a collection of 8 conversations, each consisting of six dialogue rounds, forming a complete story. Then, we recruited two graduate students in education to create another 8 stories via six rounds of human-to-human dialogue. We provided these human narrators with a detailed list of specific instructions like the list of mathematical terms they have to use. The story content told by one speaker during each dialogue round is deemed a story continuation. In total, we compiled 96 LLM-crafted and 96 human-crafted story continuations for our evaluation.

C ITEMS IN THE MATHEMATICAL LANGUAGE ASSESSMENT IN THE USER STUDY

C.1 Mathematical Language Definition

- (1) What does the word "equal" mean?
- (2) What does the word "half" mean?
- (3) What does the word "add" mean?
- (4) What does the word "subtract" mean?
- (5) What does the word "estimate" mean?
- (6) What does the word "sum" mean?

C.2 Mathematical Language Recall

- (1) What is one word that means "the same amount"? (Estimate, Equal, or Sum)
- (2) What is one word that means "a total amount"? (Equal, Add, or Sum)

- (3) What is one word that means “plus”?
(*Add, Subtract, or Half*)
- (4) What is one word that means “to take away”?
(*Sum, Subtract, or Estimate*)
- (5) What is one word that means “to cut in two”?
(*Add, Half, or Estimate*)
- (6) What is one word that means “a good guess” when you can’t count how many things there are?
(*Sum, Estimate, or Subtract*)

C.3 Mathematical Language Transfer — Posttest

- (1) Let’s say you have 5 books and your friend also has 5 books. How does the number of books you have compare to the number of books your friend has?
(*More, Half, or Equal*)
- (2) Let’s imagine you have a net filled with butterflies. There are so many butterflies that you can’t count them all. You want to guess how many butterflies are in the jar. What’s another word for guess in this context?
(*Sum, Estimate, or Add*)
- (3) Now, let’s imagine there is a cookie and three friends. Splitting the cookie would give each person one third. If there were two friends how much of the cookie would each person get?
(*Half, Equal, or Estimated*)
- (4) Let’s imagine you have 7 cookies in your basket and your friend has 10 cookies in her basket. You want to have the same number of cookies with your friend. How could you increase the number of cookies in your basket?
(*Subtract, Add, or Half*)
- (5) Let’s say you have two baseballs and three basketballs. All together, you have five balls. This is called a total. What’s another word for total in this context?
(*Sum, Estimate, or Subtraction*)
- (6) You have 6 stamps, and your friend has 4 stamps. If you want to find out how many more stamps you have than your friend, what should you do?
(*Add, Subtract, or Estimate*)

C.4 Mathematical Language Transfer — Pretest

- (1) Let’s say you have 5 toys and your friend also has 5 toys. How does the number of toys you have compare to the number of toys your friend has?
(*More, Half, or Equal*)
- (2) Let’s imagine you have a jar filled with candies. There are so many candies that you can’t count them all. How could you guess how many candies there are?
(*Sum, Estimate, or Add*)
- (3) Now, let’s imagine there is a cupcake and three friends. Splitting the cupcake would give each person one third. If there

were two friends how much of the cupcake would each person get?

(*Half, Equal, or Estimated*)

- (4) You want to measure out 10 grams of sugar but you currently have 8 grams. What action should you take to put 2 more grams of sugar onto your scale?
(*Subtract, Add, or Half*)
- (5) Let’s say you buy two oranges and three apples. Now you have five pieces of fruits. What do you call the number five in this context.
(*Sum, Estimate, or Subtraction*)
- (6) Your backpack is filled with books and is too heavy. What should you do with the books to make your backpack lighter?
(*Add, Subtract, or Estimate*)

C.5 Mathematical Language Practice

The picture of assorted fruit quantities used here is shown in Fig. 10.



Figure 10: The image of assorted fruit quantities used in the practice facet of mathematical language assessment.

- (1) Now, you are going to use the word “equal” to describe what you see in this picture.
- (2) Now, you are going to use the word “half” to describe what you see in this picture.
- (3) Now, you are going to use the word “sum” to describe what you see in this picture.
- (4) Now, you are going to use the word “add” to describe what you see in this picture.
- (5) Now, you are going to use the word “subtract” to describe what you see in this picture.
- (6) Now, you are going to use the word “estimate” to describe what you see in this picture.

D EXAMPLE PROMPTS FOR IMPLEMENTING THE MATHEMYTHS SYSTEM

Table D shows the full list of prompt templates used in implementing our system for each task.

Table 4: Prompt templates used in story element extraction, story continuation, question generation, intent classification, rephrase and hint, and encouragement generation. According to the ChatGPT API specification, the “system instruction” can give high level instructions to guide the model’s behavior throughout the conversation, while the “user input” provides the model with the prompt that the user has inputted or the information needed to complete specific tasks.

Task	Prompt Template	Few-Shot Examples
Story Element Extraction	• system instruction: I want you to extract the character, the character’s name, and the setting for a story from the user’s response. Please return the answer in JSON format. • user input: [the child’s responses of the protagonist, the protagonist’s name, and the setting]	• input: <i>A robot named Diego will be in the space.</i> output: {"CHARACTER": "ROBOT", "NAME": "DIEGO", "SETTING": "SPACE"} • input: <i>The character is a dog. His name is Tommy. He will be in the forest.</i> output: {"CHARACTER": "DOG", "NAME": "TOMMY", "SETTING": "FOREST"} • input: <i>I like coconuts. So the character is a coconut. Her name will be Samantha. She is in my home.</i> output: {"CHARACTER": "COCONUT", "NAME": "SAMANTHA", "SETTING": "HOME"}
Story Continuation	• system instruction: You’re a storytelling robot. Let’s play a joint storytelling game where we build a story together with the main character being a [character] named [name] in the [setting]. This story should revolve around the main character and a little kid named [the child’s name] reaching a certain location, attaining a certain object, or fulfilling a certain objective while conquering many obstacles along the way. To start, please craft a 2-sentence introduction to the story. From there, we will alternate turns, with each person adding more to the story. When it is your turn, only write the story content while using as many of the following words as possible: equal, sum, half, add, subtract, and estimate. Please keep your responses simple and appropriate for a young child. Please do not ask me any questions or respond with anything unrelated to the story. If I need to communicate with you in English, I will use curly brackets {like this}. Please be creative and have fun with this storytelling adventure! If you understand and are ready to begin, respond with only “yes”. • user input (if the story is continuing): [the child’s addition to the story] {First, in curly brackets, acknowledge my addition and commend me in a concise 10-word sentence. Next, continue the story by correctly using [mathematical term one] and [mathematical term two] or their variants within two distinct 15-word sentences. Lastly, explain their meanings within a 20-word sentence, grounding it firmly within the story’s context. Do not end the story.} • user input (if the story is reaching the end): [the child’s addition to the story] {First, in curly brackets, acknowledge my addition and commend me in a concise 10-word sentence. Next, end the story by correctly using [mathematical term one] and [mathematical term two] or their variants within two distinct 15-word sentences. Then, explain their meanings within a 20-word sentence, grounding it firmly within the story’s context. Lastly, summarize the story within a 30-word sentence.}	N/A

Task	Prompt Template	Few-Shot Examples
Question Generation	• system instruction: Given an unfinished story, compose one single, straightforward question to prompt a 6-year-old to expand on the story. The question should focus on the main character's next steps or feelings and motivate the child to use words like equal, sum, half, add, subtract, and estimate in their response. Do not mention "math", "mathematics", and "mathematical skills". Limit your response to 20 words. Simple future tense. Only reply with the question. • user input: [the agent's generated story content]	• input: <i>Samantha knew that the sum of challenges she would face would be great, but she was determined to succeed. She began her journey, keeping her eyes and ears open for any clues that would lead her closer to the gem.</i> output: <i>What challenges do you think Samantha will face halfway through the journey?</i> • input: <i>Jennie overheard a group of sea creatures talking about the cave's entrance, which could only be found when the sum of the three tallest coral reefs was equal to the depth of the sunken ship nearby. Determined to solve this riddle, Jennie decided to set out and add this great adventure to her life's experiences.</i> output: <i>How do you think Jennie will figure out the sum of the heights of the three tallest coral reefs?</i> • input: <i>The ancient prophecy stated that the Dragon's Heart would be hidden in a place where the sum of two particular rivers met. Michael realized that the location must equal to the intersection of the mighty Sapphire River and the mysterious Emerald River, deep within the enchanted forest.</i> output: <i>What do you think Michael will do to estimate where the two rivers meet?</i>
Intent Classification	• system instruction: As a language model, your task is to determine whether a given phrase expresses a feeling or emotion in the first person, such as uncertainty, agreement, satisfaction, or interest. Your response should be a binary output of either 1 or 0, where 1 indicates that the phrase does express a feeling or emotion in the first person, and 0 indicates that it does not. • user input: [the child's responses]	• input: <i>I don't know.</i> output: 1 • input: <i>I like it.</i> output: 1 • input: <i>I don't like it.</i> output: 1 • input: <i>Diego is happy.</i> output: 0 • input: <i>He is sad.</i> output: 0 • input: <i>I am not sure.</i> output: 1 • input: <i>Go on.</i> output: 1
Rephrase and Hint	• system instruction: You and a 6-year-old are playing a storytelling game, taking turns to contribute to a unique story. However, when the child fails to continue the story or just wants you to continue the story, your task is to encourage them to keep going by simplifying this question: [question], and give them some hints. Limit your response to 30 words. • user input: [the child's insubstantial response or no response]	N/A
Encourage-ment Generation	• system instruction: You and a 6-year-old are playing a storytelling game, taking turns to contribute to a unique story. However, Sometimes the child just says a few words, your task is to first acknowledge what he says and then encourage him to say more. Limit your response to 30 words. • user input: [the child's short response of story continuation]	N/A