

StoryPal: Supporting Young Children’s Dialogic Reading with Large Language Models

Kunlei He

University of California Irvine
Irvine, California, USA
kunlei.he@uci.edu

Joshua Munsey

University of California, Irvine
Irvine, California, USA
jmunsey@uci.edu

Aria Gastón-Panthaki

University of California, Irvine
Irvine, California, USA
agastonp@uci.edu

Maggie Zhang

University of California, Irvine
Irvine, California, USA
maggiegz@uci.edu

Dongni Zhuo

University of California, Irvine
Irvine, California, USA
dzhuo2@uci.edu

Mark Warschauer

University of California, Irvine
Irvine, California, USA
markw@uci.edu

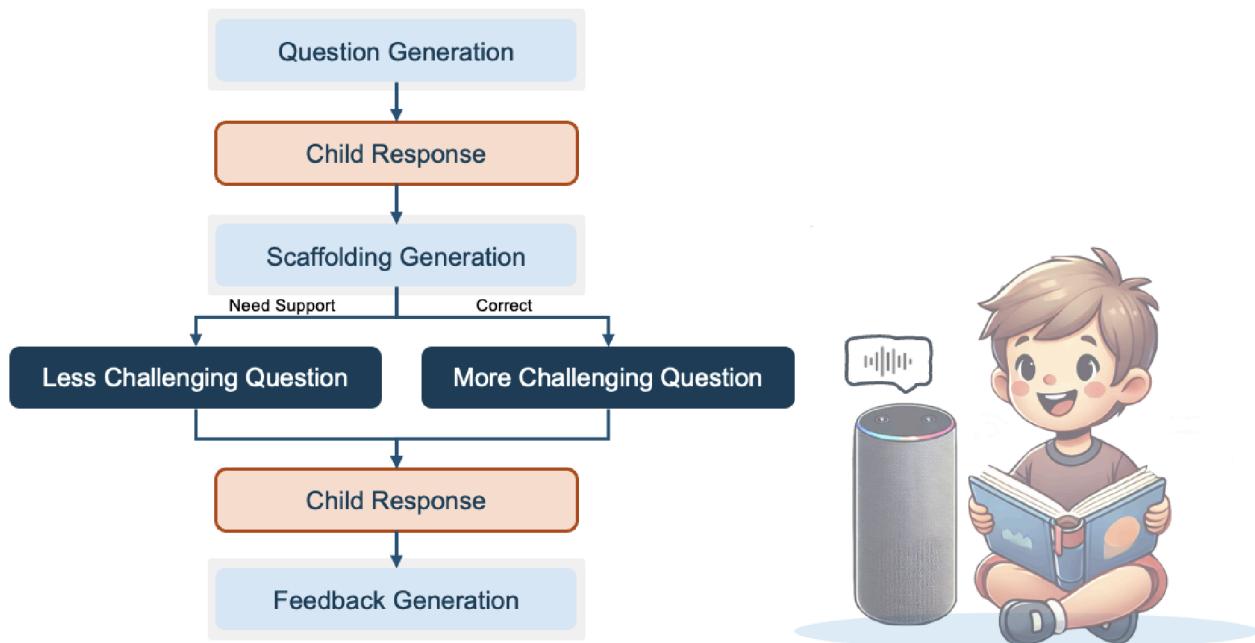


Figure 1: Interaction flow between a child and StoryPal

Abstract

Early literacy skills are crucial predictors of children’s academic success. Dialogic reading—an interactive approach where adults and children engage in discussions about stories—has proven highly effective in developing these skills. However, many families face barriers implementing this practice due to time constraints, limited resources, or linguistic challenges. We present StoryPal, an

LLM-powered conversational agent that facilitates dialogic reading through contextual questioning, adaptive scaffolding, and personalized feedback. In a study with 23 children ages 4-7 from diverse socioeconomic and linguistic backgrounds, we found high levels of verbal engagement with distinct patterns between English-dominant and bilingual children. The system’s dynamic scaffolding effectively supported struggling readers while challenging proficient ones. Parents valued StoryPal as a supplementary tool that maintained children’s reading engagement when they were unavailable, but emphasized that it should not replace parent-child interactions. Our findings demonstrate the potential of LLM-powered agents to support dialogic reading by adhering to established educational practices.



This work is licensed under a Creative Commons Attribution 4.0 International License.
IDC '25, Reykjavik, Iceland

© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1473-3/25/06
<https://doi.org/10.1145/3713043.3728867>

CCS Concepts

- Social and professional topics → Children;
- Applied computing → Education.

Keywords

Dialogic reading, large language models, young children, language development

ACM Reference Format:

Kunlei He, Aria Gastón-Panthaki, Dongni Zhuo, Joshua Munsey, Maggie Zhang, and Mark Warschauer. 2025. StoryPal: Supporting Young Children's Dialogic Reading with Large Language Models. In *Interaction Design and Children (IDC '25), June 23–26, 2025, Reykjavik, Iceland*. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3713043.3728867>

1 Introduction

Early literacy is widely recognized as a key predictor of children's long-term academic success [9, 13]. Among the various strategies for supporting early literacy, dialogic reading stands out for its interactive, conversational approach. Rather than merely reading a story aloud, caregivers or educators engage children in active discussions: asking questions, prompting them to describe scenes or characters, and encouraging them to connect the story to their own experiences [50]. This technique not only maintains children's focus and enjoyment but also stimulates language growth—boosting vocabulary, improving story comprehension, and prompting abstract thinking skills that underlie successful reading in later academic years.

However, implementing dialogic reading presents significant challenges, particularly for families with limited time, resources, or confidence in reading strategies [45]. Many caregivers, despite recognizing the importance of early literacy, struggle to set aside dedicated reading sessions or to sustain children's active engagement. In multilingual households, this challenge becomes more pronounced as children enter formal schooling: parents often desire to support children's proficiency in the school language, yet lacking both confidence and adequate resources to do so. These hurdles underscore a pressing need for accessible tools that retain dialogic reading's rich, conversational qualities while alleviating caregiver burdens and catering to varied linguistic environments.

Recent advancements in large language models (LLMs) promise scalable and potentially cost-effective ways to bring personalized dialogic reading experiences to children at home. Prior conversational agents often rely on scripted dialogue flows that do not adapt easily to each child's responses and learning progress [56]. It is also hard to scale because extensive human labor was required to write the conversation script for every single new book. Large language models (LLMs) have recently opened new avenues to overcome these challenges. These models can interpret a child's input in real time and generate contextually relevant responses, creating dynamic reading interactions that adjust to each child's understanding. Through careful prompt engineering, LLMs can be guided to follow established educational practices while tailoring their output to match individual children's needs. This adaptability eliminates the need for manually scripted conversations for each book, making it possible to scale personalized reading support across a wide range of texts and learning levels.

In this paper, we introduce StoryPal, a voice-based reading companion that harnesses an LLM to deliver personalized dialogic reading experiences. We designed StoryPal to actively engage children aged 4–7 by generating questions in real time, dynamically adjusting support based on their responses, and scaffolding children's comprehension and language production. Through a user study with 23 children from different socioeconomic and language backgrounds, we demonstrate how StoryPal can effectively maintain engagement and encourage rich verbal participation during story reading. We also highlight how families perceive this technology—its usability, benefits, and limitations—across a range of home contexts. Finally, we discuss design implications for creating LLM-powered reading tools that accommodate children's varying needs, setting the stage for more inclusive, adaptable, and scalable technologies that facilitate dialogic reading.

2 Related Work

2.1 Dialogic Reading and Early Literacy Development for School Readiness

Early literacy skills are widely recognized as a strong predictor of later academic achievement, serving as a cornerstone for educational success throughout a student's academic journey [9, 13]. Numerous longitudinal studies have shown that early reading proficiency predicts educational outcomes in the elementary and middle school years, including reading comprehension and math performance [13, 19, 49]. During the critical formative years (ages 4–7), children develop foundational reading capabilities that greatly influence their future learning: they learn to decode text, build vocabulary, gain phonological awareness, and understand narrative structures [35].

One of the most common and beneficial early literacy activities is shared story reading between the children and the caregivers. When children merely listen passively to the story, they miss opportunities to practice language skills and are more likely to become distracted and disengaged from the story. Research consistently demonstrates that young children benefit significantly more from guided, interactive reading that actively engages them in conversation about the story – an approach known as dialogic reading [50].

This approach is structured around two frameworks: **PEER** sequence (Prompt, Evaluate, Expand, Repeat) and its companion prompting guide **CROWD** (Completion prompts, Recall questions, Open-ended questions, Wh- questions, and Distancing prompts). Through the PEER sequence, adults prompt children to engage with the text using CROWD prompts, thoughtfully evaluate their responses, expand upon their contributions with additional information, and ask the child to repeat their expansion (see Table 1). While the PEER sequence provides guidelines for how to carry out back-and-forth dialogue, the CROWD framework offers specific types of prompts that engage children at various cognitive levels, from simple completion tasks to complex analytical thinking (see Table 2).

The PEER sequence and CROWD prompts in dialogic reading have consistently demonstrated their benefits in enhancing vocabulary acquisition, developing listening comprehension, and building robust narrative skills (for review, see [43]). In a seminal study,

Whitehurst et al. [50] found that children whose parents were trained in dialogic reading using these two frameworks exhibited a larger quantity of language production during reading and showed significantly higher expressive vocabulary gains compared to control groups. Similarly, in a study by Hargrave and Sénechal [23], preschool children with limited vocabularies who participated in dialogic reading sessions showed substantial gains in vocabulary knowledge compared to children who received standard reading practices, reinforcing the effectiveness of the PEER and CROWD frameworks in promoting language development.

However, despite the proven effectiveness of dialogic reading, many children face substantial barriers in accessing such quality literacy practice. Persistent socioeconomic inequalities, increasingly limited caregiver availability, and the lack of access to training programs for parents often result in reduced opportunities for interactive reading at home. Without consistent exposure to high-quality reading practices, children have fewer opportunities to develop crucial literacy skills, leading to significant language and literacy gaps [45]. These educational disparities—often starkly visible even before children enter formal schooling [41]—can persist or widen throughout their academic journey, underscoring the pressing need for innovative, scalable, and effective interventions that can bridge these gaps.

The growing prevalence of multilingual households has introduced new challenges and opportunities in fostering early literacy development. An increasing number of children worldwide grow up exposed to more than one language at home. In the US, for example, nearly one in four children live in households where a language other than English is spoken [3]. While multilingualism can provide cognitive, social, and cultural benefits [5], it also introduces unique challenges as these children start formal schooling. They need to learn how to speak, read, and write in the majority language at school (such as English in the US) while simultaneously maintaining interactions in their home language outside of school. Research demonstrates that home language skills can positively transfer to the other language [14]. At the same time, studies highlight that early bilingual exposure is crucial for the development of both languages and supports children's academic success [30]. For many bilingual families, this necessitates finding a balance between fostering the home language and creating meaningful opportunities for children to engage with the school language, particularly in preparation for formal education. Many bilingual families express a desire for additional resources to support their children's literacy development in the school language, especially as they approach school age [46]. In such contexts, parents often seek supplementary technological tools—such as mobile apps—that provide playful learning activities in the school language [39]. This need is especially pronounced among caregivers who feel less confident in the majority language themselves [22].

2.2 Conversational Agents as Children's Language Learning Partner

Recent advances in artificial intelligence have led to the development of conversational agents (CAs) that can simulate interactive, interpersonal communication in real time. These technologies offer promising potential in supporting early language and literacy

development, particularly for children facing barriers accessing consistent support for literacy activities at home. Unlike text-based chatbots, CAs carry out natural, spoken dialogue, making them especially accessible to young learners who are still acquiring reading and writing skills. By engaging children in interactive conversations, CA provides meaningful language exposure, ample practice opportunities, and instant feedback, all of which are important for early language development [47].

Conversational agents such as smart speakers (e.g., Amazon Echo, Google Home) are increasingly present in family settings, and recent research suggests growing adoption among younger users. In the United States, 88% of the children aged 3 - 12 have access to smart speakers, with half using them daily [6]. Similarly, in the UK, more than eight in ten (83%) children use smart speakers for a wide variety of tasks, including searching for information, listening to audiobooks, etc. [40]. For parents concerned about screen overexposure, the availability of such a screen-free, voice-based alternative can be especially appealing. Moreover, children often perceive smart speakers as friendly approachable conversation partners [57], which may encourage continued engagement and open up opportunities for language learning, interactive storytelling, and other educational activities without the potential drawbacks of prolonged screen use. Indeed, Xu et al. [54] demonstrated this potential by using smart speakers to conduct dialogic reading: the device narrated the story, guided page turning, posed questions based on the story, waited for the child's response, and provided feedback. Notably, children who engaged in dialogic reading with the speaker showed the same level of story comprehension as those who read with a human.

Although CAs offer promising benefits for language learning, they still present challenges in the areas of personalization and scalability. Many current CA implementations rely on pre-scripted conversations [12, 56], wherein educators or developers must manually craft prompts, questions, and feedback for each book. This process is labor-intensive and time-consuming, making it difficult to adapt the system to new content or to scale up for use with diverse reading materials. Furthermore, limited personalization constrains the system's capacity to respond to children's varied, creative inputs. Because many CAs operate on fixed response categories or heavily templated dialogue flows, they are often unable to offer tailored feedback that acknowledges and builds on each child's unique linguistic strengths, interests, or mistakes [33, 52]. This rigidity can lead to interactions that feel repetitive and unresponsive, diminishing the potential for sustained engagement and impactful learning gains. Therefore, addressing these limitations is crucial to realize the full potential of conversational agents for interactive and meaningful language-learning experiences.

2.3 Large Language Models for Children's Learning

Recent progress in large language models (LLMs) holds promise for overcoming the limitations in personalization and scalability faced by CAs for children's learning. These models exhibit sophisticated capabilities in processing natural language input, understanding context, and generating coherent, contextually appropriate responses [38]. They can also reason about abstract concepts [26],

Component	Description	Example
Prompt	Adult initiates interaction by asking questions about the story	“What do you think will happen next?”
Evaluate	Adult provides feedback on child’s response	“Yes, that’s right!” or “It might look like a dog, but it’s a wolf.”
Expand	Adult adds information to child’s response	Child: “It’s a dog” Adult: “Yes, it’s a spotted Dalmatian dog”
Repeat	Adult asks child to repeat expanded response	“Can you say ‘spotted dalmatian dog’?”

Table 1: The PEER Sequence for Dialogic Reading

Prompt Type	Description	Example
Completion	Prompts that ask the child to complete a sentence.	“The cow jumped over the ___.”
Recall	Prompts about events or characters the child has already read.	“What did the hungry caterpillar eat first?”
Open-ended	Prompts that ask the child to describe the story or express their ideas about the story.	“How do you think the bear felt after receiving the gift from the girl?”
Wh-question	Prompts that start with who, what, where, when, why.	“Where did the family go for their picnic?”
Distancing	Prompts that connects story to the child’s personal experience	“Have you ever felt scared like the little bear?”

Table 2: The CROWD Prompts for Dialogic Reading

follow multi-step instructions [15], and adapt their communication style based on the user’s age and comprehension level [36]. Through prompt engineering [11]—a method of providing specific instructions and examples to guide model behavior—developers can leverage pre-trained LLMs without extensive model finetuning. While the performance of prompted LLMs may not always surpass models fine-tuned for specific tasks, this approach offers a flexible and efficient way to utilize LLMs effectively in education [28], health care [58], programming [51], etc.

LLMs are currently being deployed in two primary approaches to support children’s learning. The first approach positions LLMs as behind-the-scenes content generators, creating educational materials that parents and teachers can use in their interactions with children. This method does not involve direct child-LLM interaction. For instance, Dietz et al. [17] demonstrated how GPT-3.5 could generate targeted comprehension questions from story books using the CROWD framework (see Table 2), facilitating meaningful parent-child discussions. In a similar vein, Lee and colleagues [32] used GPT-4 to create personalized stories incorporating new vocabulary words identified in children’s surroundings. These research efforts have shown that when LLMs are guided by carefully designed prompts that incorporate established educational principles, they can produce high-quality learning materials.

The second approach involves creating conversational agent (CA) powered by LLMs that children directly interact with. Unlike traditional CAs that are often constrained by pre-determined dialogue trees, LLM-powered CAs can engage in natural, free-flowing conversation, maintain context throughout the interactions, and respond

relevantly to unexpected inputs. Recent research has demonstrated promising applications of these child-focused CAs. For example, Zhang and colleagues [59] developed Mathemyth, a CA that leverages LLMs to engage preschoolers in collaborative oral storytelling while naturally incorporating mathematical vocabulary and concepts. Their study showed that children found these interactions engaging and learned mathematical terminology as effectively as they did through human-led storytelling. Building on the success of using LLMs to generate reading questions and facilitating child-friendly conversation, our study further extends the application of LLMs for younger children’s literacy and language development through dialogic reading.

3 The Development of StoryPal

Drawing from extensive research on dialogic reading and the advanced capabilities of LLMs, we developed StoryPal through an iterative design process. Our goal was to create an interactive reading companion tailored for children aged 4 to 7. This system serves as an exploratory investigation into how LLMs can function as reading partners for young learners. Table 3 presents a comparison between StoryPal’s LLM-based approach and traditional rule-based CAs, underscoring the key differences in functionality and implementation that informed our design decisions.

At its essence, StoryPal is designed to promote children’s verbal engagement during story reading. The system accomplishes this by dynamically generating story-relevant questions and providing individualized support and feedback based on each child’s response. This approach creates an interactive reading experience that adapts

Feature	Rule-Based CAs	StoryPal (LLM-Based)
Question Generation	Limited to pre-scripted questions with predefined answers	Dynamically generates diverse questions following CROWD framework in real-time
Response Processing	Can only recognize expected answers or limited variations	Can interpret and respond to a wide range of child utterances
Scaffolding Approach	Basic remediation for incorrect answers	Adaptive bi-directional scaffolding
Feedback Mechanism	Pre-determined feedback options	Contextually relevant feedback that builds on child's specific responses
Content Scalability	Requires manual scripting of questions for each new book	Automatically generates appropriate questions for any book without additional programming
Content Safety	Fully controlled through pre-vetted content	Requires extra safety protocols to prevent inappropriate responses

Table 3: Comparison Between StoryPal (LLM-Based) and Rule-Based Conversational Agents [12, 25, 52] for Dialogic Reading

to the child’s comprehension levels. The complete interaction flow of the StoryPal system is presented in Figure 2.

Building on recent successes in integrating Large Language Models (LLMs) into children’s educational applications [17, 59], we grounded the development of StoryPal in established educational principles, specifically those from dialogic reading research, which emphasizes how adults can engage children through meaningful questions and guide back-and-forth conversational exchanges during reading. Through iterative refinement, we translated these educational principles into effective LLM prompts. The following sections detail: 1) the design rationale that guided StoryPal’s development, 2) our approach to prompt engineering, and 3) the implementation of StoryPal.

3.1 Design Rationales

DR1: Enable Natural Back-and-Forth Reading Interactions. Our primary design rationale is to replicate the effective dialogic reading practices typically observed between caregivers and children. We build upon the PEER sequence (Prompt, Evaluate, Expand, Repeat) presented in Table 1. In a typical interaction, StoryPal first prompts the child with a question (e.g., “What do you see on this page?”), evaluates their response (e.g., “Yes, that’s right!”), expands on their answer (e.g., “And not only is it a caterpillar, it’s a very hungry caterpillar who’s looking for food”), and repeats the concept to reinforce learning (e.g., “Can you tell me again what kind of caterpillar this is?”). The adoption of the PEER sequence extends previous dialogic reading systems that typically focus only on asking questions and providing feedback [56] by incorporating the expansion and repeating phase where the system builds upon children’s responses to deepen their understanding and provide additional opportunities for practice and reinforcement.

DR2: Employ Diverse Question Types. To facilitate rich discussions about the story, StoryPal employs the CROWD questioning framework presented in Table 2, which encompasses five types of questions: Completion, Recall, Open-ended, Wh-questions, and Distancing. Each question type targets different cognitive skills and learning objectives [50]. Completion prompts (e.g., “The caterpillar

ate a red __”) help build children’s phonemic awareness and familiarize them with the structure of language (citation), while Recall questions (e.g., “Remember what the caterpillar ate on Monday?”) reinforce story comprehension and narrative sequences. More complex cognitive processes are engaged through Open-ended questions (e.g., “What do you think will happen to the caterpillar after it eats all that food?”), which foster prediction and reasoning abilities, and Wh-questions (e.g., “Why did the caterpillar get a stomachache?”), which develop analytical thinking and causal understanding. Distancing questions (e.g., “Have you ever felt really full like the caterpillar?”) help children connect the story to their personal experiences, promoting deeper engagement and real-world application of concepts. By strategically employing this range of question types, StoryPal fosters children’s development from basic comprehension to higher-order thinking skills while maintaining their engagement through different levels of cognitive challenges.

DR3: Emphasize Dynamic Scaffolding. The PEER framework’s linear sequence, while useful, does not explicitly address how to adapt when children struggle with a concept, provide incrementally more challenging questions when they succeed, or maintain engagement when they show hesitation. Scaffolding, which involves dynamically adjusting the level of support to match a child’s current abilities and gradually reducing assistance as they progress, is crucial for effective learning with conversational agents [52, 59]. This limitation of PEER led us to develop additional scaffolding strategies that complement and enhance the basic structure.

StoryPal implements an adaptive scaffolding mechanism that responds dynamically to children’s engagement and comprehension levels. For correct responses, StoryPal validates the child’s understanding and strategically increases the complexity of follow-up questions. For example, if a child correctly identifies “The caterpillar ate an apple” in response to a recall question, StoryPal might say “That’s right! The caterpillar did eat a bright red apple,” and then advance to a more challenging question such as “Why do you think the caterpillar was still hungry after eating the apple?” When children provide incorrect responses, StoryPal maintains an encouraging tone while offering graduated assistance. For instance,

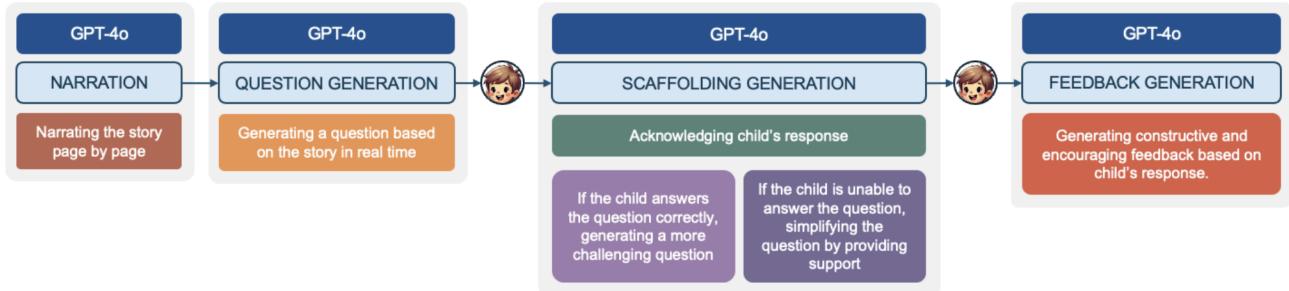


Figure 2: The interaction flow of the StoryPal system

if a child misidentifies the apple as an orange, StoryPal might respond, “That’s a thoughtful guess! Let’s look at this picture more carefully together. I see that this fruit is red and round, and it has a little green leaf on top. Is that an orange or an apple?” By providing hints and offering a simpler choice between two options, StoryPal reduces the task complexity while guiding the child toward the correct answer. For children who remain silent or show hesitation, StoryPal demonstrates patience and provides explicit encouragement: “It’s okay to take your time thinking about it. When you’re ready, can you tell me what fruit the caterpillar is eating on this page?” Through these varied scaffolding strategies, StoryPal creates a supportive learning environment that adapts to each child’s needs while gradually building their comprehension.

3.2 Prompt Engineering

StoryPal leverages GPT-4o, a leading language model. Through prompt engineering, GPT-4o performed three primary tasks aligned with the dialogic reading principles: 1) generating contextually appropriate questions to encourage children’s active engagement with stories; 2) providing adaptive scaffolding based on children’s responses; and 3) delivering encouraging feedback that builds upon children’s answers while maintaining engagement. The prompts were developed through an iterative process. We adopted Molick & Molick’s prompt engineering principles [37] and Dietz et al.’s question generation method [17] as our foundation. Through a month-long revision process involving internal evaluations with our research team and pilot testing with children, we continuously refined our prompts. The final version of prompts used in the study is detailed in the Appendix.

3.2.1 Question Generation. Following the prompt structure suggested by Molick & Molick [37], the question generation prompts begin with a clear role definition as the baseline prompt: “You are a dialogic reading assistant for a child aged four to seven. Your task is to generate one prompt for the child based on the story to help the child comprehend the story better and practice their language skills.” To ensure high-quality question generation, we provided comprehensive specifications for each question type in the CROWD framework, incorporating Dietz et al.’s [17] enhanced method that goes beyond basic CROWD definitions. For each question type, we considered their content requirements, grammatical structure, and story relevance. For example, completion questions are marked

by having a blank to be filled at the very end. Wh-questions start with question words such as who, what, when, where, why, and how, and revolve around information present on the current page being read. Recall questions, on the other hand, often span several pages, asking children to recall information that has been introduced. Moreover, we provide three examples for each type of the questions, as suggested by Brown et al. [8]. This detailed specification approach, as demonstrated by Dietz et al. [17], led to higher quality questions as rated by educators compared to using basic CROWD definitions alone.

Question Quality Evaluation. To assess our question generation approaches, we compared two prompting methods: a baseline prompt and an enhanced version incorporating the CROWD framework. The enhanced version built upon the baseline prompt by adding CROWD framework specifications, while the baseline condition used only the original basic instructions. Our dataset comprised 19 stories from Project Gutenberg, utilizing the page-separated versions by Xu et al. [55]. In total, we generated and evaluated 422 questions (211 questions per condition). We evaluated the generated questions across four key dimensions: type variety, inspirational quality, story relevance, and readability. The latter three evaluation metrics were adapted from Zhang et al. [59], who developed them to assess story-based question generation prompts for children’s math learning. For the type variety dimension, two coders independently categorized each question according to the CROWD framework to analyze the distribution of question types across conditions. This analysis revealed how the framework influenced the range of generated questions. For the remaining three dimensions (readability, story relevance, and inspirational quality), the same two coders independently evaluated the questions using a 5-point Likert scale. A detailed breakdown of the evaluation metrics is provided in Table 4.

Analysis of question types revealed that the Baseline prompt predominantly generated Wh-question (61.61%) and Open-ended questions (37.44%), with minimal representation of other question types. In contrast, the CROWD prompt produced a more balanced distribution across categories, with Wh-prompts questions being the most common (42.65%), followed by Open-ended questions (27.91%), Distancing (18.48%), and Recall questions (10.43%). The CROWD prompt also generated some Completion questions (1.42%), while the Baseline prompt produced none. Overall, the CROWD

Metric	Description	IRR
Type Variety	Distribution analysis of questions across CROWD categories, comparing the presence and frequency of different question types between baseline and experimental conditions.	0.87
Readability	The generated question is written with clear, age-appropriate language that children can follow.	0.82
Story Relevance	The generated question is relevant to the key events in the story plots.	0.89
Inspirational Quality	The generated question prompts children to think actively and express their thoughts.	0.85

Table 4: Metrics used in evaluating the Baseline and CROWD prompts in question generation. The inter-rater reliability is displayed for each metric.

prompt appears to elicit a more diverse range of question types compared to the Baseline approach, as shown in Table 5.

Regarding the quality of the questions (see Table 6), the results revealed that questions generated using both methods maintained comparable readability ($t(420) = 0.47, p = .639$) and story relevance ($t(420) = 1.78, p = .076$), suggesting both approaches produced questions that were easy to understand and effectively addressed key story elements. Most notably, questions generated using the CROWD prompting method exhibited significantly higher inspirational quality (Mean = 4.13, SD = 0.66) compared to the baseline condition (Mean = 3.50, SD = 0.80; $t(420) = -8.90, p < 0.001$). This substantial difference in inspirational quality aligns with the CROWD method’s more diverse question distribution, particularly its increased use of distancing questions and open-ended prompts that encourage deeper engagement and critical thinking. These findings suggest that the CROWD method successfully achieves its primary goal of creating more thought-provoking prompts while maintaining comparable levels of readability and story relevance.

3.3 Scaffolding Generation

Our scaffolding mechanism was designed to provide adaptive, encouraging responses through a structured three-step approach. Each scaffolding prompt instructs GPT-4o to: 1) evaluate the child’s response with encouragement, 2) expand upon their answer, and 3) pose an appropriate scaffolding question. The specific scaffolding structure varies based on the accuracy of the child’s response.

For correct answers, the model first offers specific praise (e.g., “Good thinking, you got it right!”), then briefly expands the child’s response by adding story-relevant information, and finally poses one of two types of follow-up questions: either exploring other aspects of the story or connecting the story to the child’s personal experiences. For example, after a correct response about a character’s emotions, the model might ask “Have you ever felt upset before? What happened?” For incorrect or unclear responses, the model provides necessary support: it first acknowledges the child’s effort, then provides a targeted hint without revealing the answer, and finally rephrases the original question in a simpler format (using multiple choice or yes-no questions) to give children another opportunity to respond. For instance: “That’s an interesting idea. The wolf is pretending to be someone the girl knows. Which person is the wolf pretending to be, the girl’s friend or grandma?”

3.4 Feedback Generation

To conclude each interaction turn, the model replied to the child’s response to the follow-up question by providing constructive closure while maintaining children’s engagement and confidence. The feedback prompt instructs GPT-4o to deliver brief, age-appropriate responses without introducing new questions, focusing instead on acknowledging the child’s contribution and providing gentle guidance when needed. For correct responses, the model offers simple, specific praise that reinforces the child’s understanding (e.g., “You’re right. The wolf is pretending to be the girl’s grandma”). For incorrect responses or non-responses, the model provides constructive feedback that maintains sensitivity to the child’s efforts while subtly incorporating the correct information (e.g., “Hmm, pretending to be a dog might work, but the wolf is pretending to be the girl’s grandma”).

3.5 An Iterative Process

The development of StoryPal’s prompts followed an iterative process to ensure their effectiveness. Initially, the research team, consisting of graduate students who specialized in children’s language development, conducted internal testing and evaluations to refine the structure and content of the prompts. This was followed by pilot testing with five children, after which the prompts were further adjusted to ensure age-appropriateness and alignment with dialogic reading principles. Based on observations and feedback from the pilot tests, the team refined the examples for each prompt type and improved the scaffolding instructions for better adaptability to children’s responses. This continuous cycle of evaluation and revision ensured that the final prompts were both robust and responsive to the needs of young learners.

3.6 System Implementation

StoryPal is a system designed to operate on Windows platforms, requiring only a connected speaker for audio interaction. The system functions without the need for a graphical user interface, leveraging OpenAI’s latest GPT-4o-audio-preview model¹ to process both text and audio with low latency. Unlike traditional systems that rely on separate automatic speech recognition (ASR) and text-to-speech (TTS) components, StoryPal directly sends children’s voice input to

¹<https://platform.openai.com/docs/models/gpt-4o-audio-preview>

Type of questions	Baseline Prompt	CROWD Prompt
Completion	0	3 (1.42%)
Recall	0	22 (10.43%)
Open-ended	79 (37.44%)	57 (27.91%)
Wh-question	130 (61.61%)	90 (42.65%)
Distancing	2 (0.95%)	39 (18.48%)

Table 5: Distribution of Question Types Generated by Baseline and CROWD Prompting Methods

Metric	Baseline Prompt	CROWD Prompt	T-test
Readability	3.96 (0.59)	3.93 (0.67)	$t(420) = 0.47, p = .639$
Story Relevance	4.19 (0.58)	4.09 (0.61)	$t(420) = 1.78, p = .076$
Inspirational Quality	3.50 (0.80)	4.13 (0.66)	$t(420) = -8.90, p < .001^{***}$

Table 6: Quality Evaluation of the questions generated by the Baseline and CROWD prompt. Standard deviation in parentheses. Statistically significant results are reported as $p < 0.05^*$, $p < 0.01^{}$, $p < 0.001^{***}$**

the API, which generates real-time audio responses without intermediate transcription. The system manages the conversation flow through three primary modules: the question generation module, the scaffolding generation module, and the feedback generation module. Prompts for each module are provided to the API as text inputs, as presented in the Appendix. Each time after the agent asks a question, the microphone of the smarter speaker is activated to capture the child’s verbal response. Children’s voice input is then recorded and sent to the API, which interprets the child’s responses in real-time and generates appropriate conversational replies in audio format. This real-time processing ensures a fluid and responsive dialogue, keeping children engaged throughout the reading session. Prior to reading with StoryPal, users need to upload the story texts of their choice. During reading sessions, StoryPal narrates the story through the speaker, generating questions, scaffolding, and feedback on-the-fly, and listens for verbal responses from children, facilitating a voice-based interactive experience. By combining a straightforward technical setup with an advanced language model, StoryPal provides a seamless and natural reading experience, particularly for young children interacting with their printed books.

4 User Study

To evaluate the usability of StoryPal, we conducted a user study involving 23 children who engaged in a dialogic reading session using the system. The goal was to explore both the interaction dynamics and perceptions of StoryPal, addressing the following research questions:

- RQ1: What types of questions does StoryPal generate?
- RQ2: How do children verbally engage with StoryPal during dialogic reading?
- RQ3: How do children and parents perceive StoryPal as a reading companion?

**Figure 3: Child Participants reading with the agent. Left: individual child reading. Right: siblings reading together**

4.1 Participants

We recruited 23 children from 4 to 7 and their parents for the user study from two sites. The first site was a non-profit community organization serving predominantly low-income Latine families, from which 11 children and their parents were recruited. These child participants were all bilingual in Spanish and English, with Spanish being the primary language spoken at home. The second recruitment site was a participant database maintained by a public university in the southwestern United States, which yielded 12 children and their parents from middle-class households, including three pairs of siblings. These child participants were primarily English-dominant and used English as their primary language at home. Parents were included in the study to observe their children’s technology and participated in the interviews. Detailed participant information, including family income, reading frequency in English, and prior use of conversational agents (CA), are presented in Table 7. Each parent-child pair was compensated with \$40 along with a picture book for their time.

4.2 Procedures

User study sessions were conducted in different locations based on the recruitment site: Latine families participated at their local community center, while families recruited through the university database attended sessions in the university laboratory. While the

ID	Site	Age	Gender	Ethnicity	Family Income	Home Lang.	Reading Freq. in Eng.	Freq. of Using CA
							Days/week, Time/day	
1	CC	7	Male	Hispanic	30001-60000	Spanish	1 day a week, 0-10 min	<once a month
2	CC	7	Male	Hispanic	60001-90000	Spanish	5 days a week, 20-30 min	Never
3	CC	6	Female	Hispanic	30001-60000	Spanish	0 days a week, 0-10 min	Never
5	CC	7	Female	Hispanic	90001-120000	Spanish	0 days a week, 0-10 min	Never
6	CC	6	Male	Hispanic	0-30000	Spanish	5 days a week, 20-30 min	Never
8	CC	6	Male	Hispanic	0-30000	Spanish	3 days a week, >30min	Never
9	CC	7	Female	Hispanic	0-30000	Spanish	2 days a week, 0-10 min	Never
15	CC	7	Male	Hispanic	0-30000	Spanish	0 days a week, 0-10 min	<once a month
18	CC	7	Male	Hispanic	0-30000	Spanish	1 day a week, 0-10 min	Daily
19	CC	7	Male	Hispanic	0-30000	Spanish	3 days a week, 10-20 min	<once a month
12	CC	7	Male	Hispanic	30001-60000	Spanish	7 days a week, 10-20 min	Weekly
241	Lab	5	Female	White	>120001	English	7 days a week, 20-30 min	Daily
240	Lab	5	Male	Asian	>120001	English	5 days a week, 0-10 min	<once a month
252*	Lab	5	Female	Black	>120001	English	7 days a week, 20-30 min	Daily
253*	Lab	4	Male	Black	>120001	English	7 days a week, 20-30 min	Daily
254	Lab	6	Male	Hispanic	90001-120000	English	5 days a week, 0-10 min	Daily
264*	Lab	6	Female	White	60001-90000	English	7 days a week, 20-30 min	Daily
265*	Lab	4	Male	White	60001-90000	English	7 days a week, 20-30 min	Daily
272	Lab	7	Male	White	60001-90000	English	6 days a week, 20-30 min	Weekly
211	Lab	7	Female	White	>120001	English	7 days a week, 20-30 min	Daily
262	Lab	5	Male	White	>120001	English	7 days a week, 20-30 min	Never
98*	Lab	7	Female	Hispanic	>120001	English	7 days a week, >30 min	Daily
99*	Lab	6	Female	Hispanic	>120001	English	7 days a week, >30 min	Daily

Table 7: Demographics of child participants. CC stands for the community center serving predominantly Latinx families. Asterisks indicate siblings in consecutive children.

system was designed for individual use, as an exploratory component, we had sibling pairs read together to understand the dynamics of shared interaction with the system. For the study, we selected *The Impossible Mountains* by New York Times bestselling artist David Soman, a book rated 4.2 on GoodReads with content suitable for children aged 4 to 7. Each session began with a warm-up conversation where children discussed their feelings and favorite colors with the AI agent. This was followed by a 15-minute reading session (see Figure 3). After the reading activity, two concurrent interviews took place: the lead researcher interviewed the child about their perceptions while a second researcher, who was a native Spanish speaker, conducted the parent interview. Latinx parents were given the choice to conduct their interview in either Spanish or English, with 6 out of 11 parents choosing to communicate in Spanish.

4.3 Measures

Our analysis focuses on children’s verbal and visual engagement with and perceptions of StoryPal, as well as parents’ opinions on it. For sibling pairs, each child was analyzed as an independent unit for the quantitative analysis while qualitative observations were used to capture their interaction dynamics.

4.3.1 Types of Questions Generated by StoryPal. We used GPT-4 to develop a question classifier for questions generated by our system. The classifier employed prompts similar to those used in question generation. Due to our scaffolding approach, which converted initial questions into multiple-choice formats when children answered incorrectly, we added a multiple-choice category alongside CROWD classifications. Following automated classification, the first author reviewed and revised the results as needed. The classifier achieved an inter-rater reliability of 0.89 with human ratings.

4.3.2 Children’s Verbal Engagement with StoryPal. Verbal engagement was assessed through transcribed reading sessions, where research assistants recorded children’s responses verbatim. Each child response served as a distinct coding unit and was analyzed across four dimensions. We used the following metrics to evaluate children’s verbal interaction patterns, which have been validated in prior research examining children’s engagement with conversational agents [25, 52]: response rate, accuracy, relevance, and length. **Response rate** was measured as the frequency of verbal responses to questions. **Response accuracy** was coded dichotomously (0 or 1). For open-ended and distancing questions, which sometimes do not have a definitive answer, if a child said something considered

relevant to the question, we coded it as 1, otherwise 0. **Response length** was quantified by word count. **Siblings' co-reading dynamics** during the reading sessions qualitatively, analyzing patterns of turn-taking, elaboration on each other's responses, and instances of collaborative meaning-making through their dialogue.

4.3.3 Children's Perception of StoryPal. We assessed children's attitudes toward StoryPal through a structured questionnaire comprising eight items that examined perceived enjoyment, friendliness, and intelligence of the system. The protocol, adapted from Dietz et al. [16], implemented a graduated response format. Children initially indicated their basic stance (yes/no) on each item before specifying the intensity of their response ("a bit" versus "definitely"). This two-tiered approach yielded responses that were subsequently quantified on a 4-point scale (1 = "definitely no" to 4 = "definitely yes"). To enrich our understanding of children's ratings, we also prompted them to articulate their reasoning for each response.

4.3.4 Parents' Perception of StoryPal. Parent perceptions were gathered through 15-minute semi-structured interviews, which were audio-recorded and transcribed verbatim for analysis. Following Braun and Clarke's [7] thematic analysis procedures, the first three authors collaboratively developed initial broad categories: existing reading habits, system design and performance, child learning and engagement, and usage context and integration. The researchers began by jointly analyzing two English interview transcripts and established an intraclass correlation coefficient of 0.80. Subsequently, the second author, a native Spanish speaker, analyzed all six Spanish interview transcripts, while the first and third authors divided the remaining English transcripts between them. The researchers then collectively reviewed quotes within each broad category to develop more nuanced subcodes. Through this process, we identified key themes in parents' perspectives on StoryPal's integration into their children's reading practices and its perceived benefits and limitations.

4.4 Findings

4.4.1 RQ1: What types of questions does StoryPal generate? Wh-questions that probe comprehension of key events in the story were the predominant question type generated by the dialogic reading agent, comprising 58.33% (196) of all questions. Open-ended questions, which require children's own description or interpretation of the story, made up 14.58% (49), while distancing questions that connect the story to children's life accounted for 19.35% (65). Multiple-choice questions were less common at 5.95% (20), and recall questions were minimal at 1.79% (6). There were no completion questions generated by the agent. The absence of completion questions reflects the book's non-rhyming, non-repetitive nature [50]. This aligns with Dietz et al.'s findings, where LLM-generated discussion questions for shared reading similarly yielded minimal completion questions [17]. Table 8 presents the detailed distribution of different types of questions.

Notably, there was a shift in question distribution between initial and followup questions, with Wh-questions decreasing from 85.63% to 33.52%, while distancing questions increased from 0.63% to 26.26%. Multiple-choice questions only appeared as follow-up questions. This distribution of question types across initial and

follow-up interactions aligns well with established models of story comprehension. The heavy emphasis on Wh-questions during initial interactions (85.63%) helped establish basic story elements and narrative understanding – the fundamental "who, what, where, when, why" components that form the foundation of comprehension.

The shift toward more Distancing questions (26.26%) and Open-ended questions (18.18%) in follow-up interactions reflects a scaffolded approach to deeper comprehension. This transition encouraged children to move beyond literal understanding to make personal connections and engage in higher-order thinking, which is a crucial step in developing advanced comprehension and language skills. The strategic introduction of Multiple-choice questions (11.36%) as a scaffolding tool demonstrated an adaptive approach to supporting children who may need additional structure to bridge the gap between literal and inferential understanding. Table 9 shows examples of how questions progressed from the initial to follow-up level. This questioning pattern effectively supports the hierarchical nature of story comprehension, where basic understanding serves as a foundation for more sophisticated interpretation and personal connection – a progression that aligns with influential theories of reading comprehension development.

4.4.2 RQ2: How do children verbally engage with StoryPal during dialogic reading?

Overall Verbal Engagement Patterns. Children showed consistently high engagement with StoryPal throughout their reading sessions, responding to 74.73% of all questions. Their participation varied by question type. While they responded to 90.90% of multiple-choice questions, they were more hesitant with recall questions, answering only 33.33% of them. Other question types maintained steady response rates averaging 74.36%. The high response rate for multiple-choice questions likely reflects the structured nature of these prompts, which reduced cognitive load by providing clear options for children to choose from. In contrast, the lower engagement with recall questions may stem from their increased cognitive demands, as children needed to remember and retrieve specific story details from earlier pages. The consistent response rates for other question types, aligning with previous research [52], suggest that children were generally comfortable engaging with questions generated by StoryPal.

The accuracy of children's responses (60.32% overall) suggests children were appropriately challenged by the questions. This level of performance aligns with educational research showing that 50% - 70% accuracy provides an ideal balance for learning and assessment [21]. When tasks are too easy (resulting in very high accuracy) or too difficult (resulting in very low accuracy), they tend to be less effective for learning. A moderate level of challenge ensures children remain engaged while having sufficient opportunities to learn from both successes and mistakes. Children performed differently across question types. They struggled most with recall questions (33.33% accuracy) but showed strong performance on open-ended questions (69.59% accuracy). These patterns suggest that while children found memory-based tasks challenging, they were more confident expressing their thoughts when given flexibility in their responses.

In terms of response length, children produced utterances averaging 4.47 words, which aligns with findings from previous studies

	Initial	Follow-up	Overall
Completion	0	0	0
Recall	5 (3.13%)	1 (0.57%)	6 (1.79%)
Open-ended	17 (10.63%)	32 (18.18%)	49 (14.58%)
Wh-question	137 (85.63%)	59 (33.52%)	196 (58.33%)
Distancing	1 (0.63%)	64 (26.26%)	65 (19.35%)
Multiple-choice	0	20 (11.36%)	20 (5.95%)

Table 8: Question type distribution across initial and follow-up levels

Initial Question	Follow-up Question
What did Anna do to help Finn cross the river? [Wh-question]	Have you ever had to help a friend or family member with something tricky? What did you do? [Distancing]
What did Anna and Finn find while they were walking in the snow? [Wh-question]	How do you think Anna felt when she saw the bear? [Open-ended]
What did Finn do to help Anna scare the wolves away? [Wh-question]	Finn helped Anna by using something yellow to make himself look bigger. Was it a blanket or a scarf? [Multiple-choice]

Table 9: Examples of the progression of questions from the initial to follow-up level

of children interacting with conversational agents [52]. Distancing questions and recall questions elicited particularly detailed responses, with averages of 4.99 and 5.33 words respectively. The longer responses to distancing questions likely emerged as children drew connections between the story and their personal experiences, while recall questions demanded more elaborate answers as children worked to reconstruct and articulate previously encountered story details.

Verbal Engagement Patterns under Adaptive Scaffolding. StoryPal’s scaffolding mechanism was designed to adapt questions based on each child’s performance: if a child excelled, more challenging questions were presented; if a child struggled or initially remained silent, hints and simpler questions were provided. To evaluate the impact of this approach on verbal engagement, we considered two main factors: whether children answered the initial question correctly or incorrectly, and whether they participated immediately or remained silent at first, and examined how children’s verbal engagement with the follow-up questions changed accordingly.

Children showed remarkable resilience and growth when given scaffolding. Those who initially failed to answer the question correctly were able to improve their performance dramatically, with 64.58% succeeding after receiving helpful hints or simpler question options. A particularly illustrative case involved participant Child 8. When first asked “*What does Anna see when she looks down from the mountain?*”, the child remained silent, showing hesitation. However, when the question was restructured with gentle prompting—“*It’s perfectly okay if you are still thinking. Remember, Anne sees something from the mountain. Is it her village or a forest below?*”—Child 8 was able to correctly identify that Anna saw “*the village*.” This example highlights how breaking down complex questions into

simpler components and offering reassurance can help children overcome initial uncertainty and demonstrate their understanding of the material.

On the other hand, Children who started strong received productive challenges, as evidenced by their accuracy decreased to 53.74% when presented with increasingly complex questions. This decline reflected the intentional scaffolding strategy of stretching students beyond their initial comfort zone. For instance, When asked “*What did Anna and Finn decide to do when they met people at the village gate?*”, Child 3 correctly responded that they wanted to “go to the mountain.” StoryPal then deepened the inquiry, asking “*That’s right, Anna and Finn went to climb the mountain. They both were very brave. Do you remember who tried to stop them?*” At this more challenging level, Child 3 showed signs of uncertainty, making a “*hmm*” sound and beginning with “*it was...*” before ultimately being unable to complete the answer. StoryPal responded supportively: “*I can see you are thinking. The blacksmith, the mean boy, and the baker tried to stop Anna and Finn.*” Child 3’s nodding response to this feedback suggested engagement with the additional story details. Together, these performance patterns, combined with the previously discussed distribution of question types, provide strong evidence that our scaffolding mechanism successfully adapted to individual children’s needs.

Verbal Engagement Patterns by Language Backgrounds. Recall that our sample contained children with two distinct language backgrounds-half of the children were English dominant, whereas the other half were Spanish-English bilingual, most of whom spoke Spanish at home. Although children exhibited an overall high level of verbal engagement, a more fine-grained analysis revealed different patterns between the two groups.

English-dominant children showed higher overall response rates (79.16%) compared to their bilingual peers (69.87%), though accuracy levels remained comparable between the groups (61.98% and 58.52% respectively). This pattern suggests that while bilingual children may have been more selective about when to respond, they were equally capable of comprehending and correctly answering questions when they chose to participate. The most striking disparities emerged in recall and distancing questions. For recall questions, English-dominant children's response rate (50.00%) doubled that of bilingual children (25.00%). This difference was exemplified when children were asked to recount earlier story events. For example, when asked “*Can you remember any other times when Anna and Finn faced challenges in their journey?*”, Child 286 (English dominant) answered “*They tried to hop on ... hop on the rocks on the river. They tried to scare the wolves*”, while Child 12 (bilingual) stayed silent during the entire response window. This pattern suggests that recall questions posed particular challenges for bilingual children, who needed to simultaneously manage story details in working memory while constructing responses in English in the very limited time.

Analysis of response lengths revealed marked differences between English-dominant and bilingual children. English-dominant children produced longer responses, averaging 5.78 words per response with considerable variation ($SD = 7.18$, range: 2.25-27.56 words). In contrast, bilingual children's responses were shorter and showed less variation, averaging 3.48 words ($SD = 1.56$, range: 1.31-6.19 words). This disparity was most evident in two types of questions: recall and distancing. For recall questions, English-dominant children averaged 9.50 words, while bilingual children averaged 5.25 words. The gap widened further with distancing questions, where English-dominant children produced 6.54 words on average, compared to bilingual children's 2.86 words. Table 10 present two comparative examples. The data suggest that as questions demanded more sophisticated cognitive processing—such as recalling multiple events or making personal connections—the additional cognitive load of expressing complex thoughts in a less familiar language appeared to limit bilingual children's verbal responses.

Sibling Co-reading Dynamics. Although our agent was initially designed for individual use, we explored its application with three pairs of siblings who read together. Two of the sibling pairs demonstrated natural turn-taking, consecutively responding to each question. Notably, in these pairs, the second child would repeat or expand on the first child's response, which the agent processed as unified input to provide appropriate guidance and constructive feedback (see the excerpt below). However, this approach could face challenges if siblings were to offer conflicting or different answers, indicating the need for more advanced response-handling mechanisms for sibling co-use.

StoryPal: *How do Anna and Finn feel about climbing the mountain?*

Child 98: *They feel excited. They feel excited.... and brave.*

Child 99: *They feel excited.... and brave. They feel like they really want to do this.*

StoryPal: *Great you noticed their bravery and excitement. Even when the villagers tried to scare them, Anna and Finn were determined to climb the mountain. Have*

you ever felt brave to try something new like Anna and Finn?

Child 98: *I try to feel brave when I go on big roller coasters and I try to feel brave when they have dark and scary parts with loud noises.*

Child 99: *And I try to be as brave as I can.*

Story Pal: *That sounds really brave. Being on big roller coasters with dark and loud parts can be scary, but you try to be brave just like Anna and Finn.*

The third sibling pair presented a different dynamic due to their two-year age difference: a six-year-old sister and her four-year-old brother. In this case, the older sibling took on a leadership role, encouraging her brother to participate by prompting him to take his turns. Despite her efforts, the younger sibling engaged minimally during the session, which their parents attributed to his early stage of language development (see the excerpt below).

StoryPal: *Why did the wolves run away when Finn used the horn?*

Child 264: *[whispering to the brother] It's your turn now.*

Child 265: *[looking at his sister, then the book, silent]*

StoryPal: *I didn't catch you. Did the wolves run away because they were scared or happy?*

Child 265: *[looking at his sister, silent]*

Child 264: *They were scared.*

Story Pal: *Great job, remember when Anna saw the bear, she felt scared and offered him an apple. You're doing well with understanding the story.*

4.4.3 RQ3: How do children and parents perceive StoryPal as a reading companion?

Child Perception. Overall, children reported enjoying their reading experience with StoryPal ($M = 3.36$, $SD = 0.86$). When asked why, twelve children specifically mentioned StoryPal's conversational capabilities. For instance, Child 1 and Child 253 stated they enjoyed “*talking to her.*” Child 2 explained, “*When I got something wrong, she told me the answer,*” and Child 98 noted, “*I liked that she answered us and said we did a good job.*”

Children also perceived StoryPal as very friendly ($M = 3.88$, $SD = 0.34$). According to Child 1, StoryPal's friendliness stemmed from the fact that “*she never says any rude stuff.*” Likewise, Child 2 and Child 254 remarked that StoryPal was “*nice*” to them. One child (Child 240), however, felt StoryPal was not friendly because “*she doesn't hear me.*” Upon reviewing video recordings, we found that Child 240 spoke very quietly for some questions, which prevented StoryPal from registering the child's responses.

Regarding StoryPal's perceived intelligence, children on average believed StoryPal could hear ($M = 3.54$, $SD = 0.78$) and understand ($M = 3.63$, $SD = 0.88$) them. When asked why they thought StoryPal understood them, children offered varied explanations. Child 3 said, “*Because she is almost like a human, but he is a speaker,*” recognizing its nature as an intelligent device. Similarly, Child 18 explained, “*There are wires and wires connect them to the computer,*”

English-dominant Child	Spanish-English Bilingual Child
<p>StoryPal: <i>Have you ever needed help from someone like Anna helped Finn? What happened?</i></p> <p>Child 99: <i>Sometimes I need help with certain math problems cuz I think I get them right, but then at one part of the problem something something goes wrong, so I sometimes need help to figure out what goes wrong and then I fix it.</i></p>	<p>StoryPal: <i>Have you ever seen a bear in a zoo or in the wild? What did it look like?</i></p> <p>Child 3: <i>Yes, brown.</i></p>

Table 10: Examples of a English-dominant child and a Spanish-English bilingual child responding to distancing questions

and Child 254 simply stated, “*It was programmed.*” These explanations reveal how children conceptualize StoryPal’s intelligence by blending human-like qualities with their understanding of its underlying technological mechanisms, aligning with prior research on children’s perception of conversational agents [57].

Parent Perception. Parents valued how StoryPal’s interactive questioning feature enhanced children’s reading comprehension and engagement. They noted that the strategic placement of questions throughout the story, rather than just at the end, helped maintain children’s attention and prevented them from “zoning out” (Parent 98). Parent 8 also highlighted the timing of the questions, stating that “the pausing and the questions … helps the child remember and assimilate the story, and connects with it.” Several parents emphasized how this approach particularly benefited young readers still developing their reading skills, with one parent (Parent 12) mentioning its value for her children with reading difficulties. Parent 264 thought StoryPal provided “*children who aren’t reading yet a way to still participate in hearing stories.*” Spanish-speaking parents recognized the tool’s educational value in children’s English development. Parent 3 commented: “*como el mío se desarrolló hablar los dos idiomas, porque yo si no sé hablar el inglés y eso como a los niños que están aprendiendo el inglés les ayudaría mucho*” (“since mine developed speaking both languages, because I don’t know how to speak English, this would help children who are learning English a lot”).

While parents valued StoryPal as a supplementary resource when they were unavailable, they also emphasized that it should not replace parent-child reading interactions. Parent 264 appreciated how it enables children to “*still participate in hearing stories, especially if the parents aren’t available,*” and Parent 12 noted it helps keep children focused on reading when parents can’t be present. Parent 5, a parent of multiple children, highlighted its practical value: “*I can read with the little one and the AI can take care of the other two… Cuz the youngest one we need. She needs more help.*” This sentiment was echoed by Parent 8, who noted “*Lo usaría al momento de que si yo tengo que hacer algo que por ejemplo, no sé si estoy en la cocina*” (“I would use it when I have to do something, for example, if I’m in the kitchen”). However, parents also expressed caution about potential overuse, with Parent 15 warning “*se pierde la interacción tanto con los padres, los hijos*” (“the interaction between parents and children is lost”), and P211 emphasizing that “*you can’t replace, you know, like, the influence of reading with your mom or dad.*”

Parents suggested various personalization features to further improve StoryPal. They emphasized the importance of adaptability, with Parent 5 noting that “*not one size fits all*” when it comes to grade levels and book difficulty. The ability to customize features based on children’s moods and engagement levels was suggested by Parent 252, who observed that sometimes children want to read without interruption, while at other times they’re more receptive to interactive questioning. Some parents also requested specific customization options, such as Parent 211’s suggestion for “*adjustable voice options*” to accommodate different preferences.

5 Discussion

5.1 Leveraging LLMs for Open-Ended Questions and Multi-Level Scaffolding in Dialogic Reading

A central innovation of StoryPal is its capacity to generate—and respond to—open-ended and cognitively demanding prompts in real time. Previous conversational agents for dialogic reading often rely on comprehension questions with predefined answers [12, 52, 54]. While such agents can effectively check a child’s literal understanding of a story, they rarely delve into prompts that invite abstract thinking, personal connections, or imaginative speculation, as children’s responses to these highly open-ended questions were less predictable and thus more challenging for traditional rule-based dialogue systems to capture. However, evidence that compared with those literal questions, such inferential questions are more beneficial for children’s oral language development because they can elicit more complex vocabulary, grammar, and narratives from children [34, 43]. Our LLM-based approach overcomes prior constraints by dynamically handling the wide range of replies that children provide. As a result, StoryPal can routinely pose queries asking children to infer the character’s feelings or connecting the story to their own life. Indeed, parents in our study observed that such prompts kept children “*thinking about the story*” rather than merely recalling surface-level details, underscoring the value of more open-ended dialogue in dialogic reading.

Equally important, StoryPal’s scaffolding mechanism ensures that children can receive support tailored to their current progress, fostering continued learning. Many prior dialogic reading systems have focused primarily on remediation, offering additional assistance only when children give incorrect or off-topic responses [52, 56]. Because these systems rely on fixed question templates

or rigid decision trees, they rarely scaffold upward for children who have already mastered the current task. By contrast, StoryPal's adaptive LLM-based design not only detects when a child needs simpler prompts—such as multiple-choice questions or step-by-step hints—but also recognizes when a child is ready for more cognitively demanding tasks. This dual approach aligns with Vygotsky's Zone of Proximal Development [48], in which learners make optimal progress when challenges lie just beyond their comfort zone yet remain attainable with adequate support. Indeed, we observed that children who initially responded incorrectly produced increasingly accurate, elaborated answers after receiving more explicit guidance, whereas those who answered confidently were presented with deeper inferential or evaluative prompts. This real-time, multi-level scaffolding enables StoryPal to keep each child engaged at an appropriate difficulty level, thereby maximizing both comprehension and motivation.

5.2 Meeting the Learning Needs of Diverse Users using LLMs

Consistent with prior research [52], we found that children readily engaged with a voice-based agent, demonstrating high verbal response rates. However, English-dominant children responded more frequently and with more extended utterances compared to their Spanish-English bilingual peers, particularly when faced with cognitively demanding recall or distancing questions. This pattern highlights how bilingual children—who often navigate additional cognitive and linguistic demands—may produce more concise or hesitant responses, even when they fully comprehend the story content. Although LLM-based scaffolding helped bridge this gap (e.g., through hints and simplified prompts), bilingual children were still more likely to opt out of complex recall questions. Future work could explore more nuanced language support, such as incorporating code-switching cues when children struggle to formulate responses in English [44]. Such features would ensure children have an immersive English reading experience while being able to leverage their full linguistic repertoire in Spanish. For instance, Xu et al. [53] and He et al. [25] explicitly informed the children that the conversational agent was bilingual and understood both English and Spanish, allowing children to respond confidently using either language or a combination of both. Given that LLMs are capable of transcribing and translating multilingual input [60], future designs should emphasize these bilingual capabilities of the conversational agents to bilingual children to support their learning and interaction.

Beyond differences in language backgrounds, many parents expressed strong interest in personalizing interactions to each child's unique needs—such as adjusting the book and question content to their age, reading level, and interests—rather than adhering to a “one size fits all” interaction paradigm. For instance, Park et al. [42] demonstrated this potential by developing an LLM-powered chatbot for language learning that integrated students' learning preferences and prior assessment outcomes into its prompts, enabling adaptive interactions tailored to individual needs. Future research could extend this approach by embedding children's characteristics into the prompts for the agents, incorporating factors such as reading comprehension levels, topic interests, attention

patterns, and preferred questioning styles. This approach would allow the agent to adjust its story selection, scaffolding strategies, and question complexity for each individual child.

Moreover, supporting sibling co-use emerged as another area for improvement. Just as He et al. [25] found in their home study of a conversational agent that facilitates reading, it was not uncommon for siblings to use the technology together. While some sibling pairs effectively took turns and elaborated on each other's comments, others faced challenges when younger siblings needed more support. Future work could explore designing explicit supportive features for multi-user interactions, drawing from successful examples such Chan et al. [10] which used a combination of haptic and visual cues. Additionally, we could plug sibling's information separately into the system so that the system can dynamically adjust its interaction style and content difficulty based on each child's developmental level and learning progress. These enhancements could help maximize the benefits of sibling co-use while minimizing friction points that arise from developmental differences.

5.3 Ethical and Pedagogical Considerations

While most parents in our study valued StoryPal as a supplementary reading companion for their children, several voiced concerns about the possibility of StoryPal replacing traditional parent-child reading interactions. This raises important ethical and pedagogical considerations regarding the use of AI in early literacy development. As an initial exploration of generative AI's affordances and usability, the current study primarily focused on understanding how StoryPal supports children's engagement and comprehension. However, we recognize the value of parental involvement in early literacy experiences [4, 18]. Indeed, our team has previously explored designing conversational agents (CAs) that engage parents by providing discussion prompts and encouraging joint reading interactions [24, 25, 53]. These approaches have demonstrated the potential to enhance parent-child conversations, reinforcing comprehension and fostering meaningful connections around stories. To further mitigate concerns of AI replacing parental roles, future iterations of StoryPal could integrate features that explicitly promote co-reading experiences. For example, incorporating prompts that suggest collaborative storytelling activities or inviting parents to elaborate on their child's responses can strengthen family engagement. Additionally, providing parents with insights into their child's progress can facilitate deeper discussions and reinforce the learning experience.

Moreover, the responsible use of AI in children's educational settings requires careful attention to privacy, safety, and content quality. Regarding privacy, StoryPal used GPT-4o provided by OpenAI, which protects user privacy by not storing and using data passed through APIs for model training². Despite this, there is a need for dedicated regulations for generative AI products in educational contexts. Current regulations such as the U.S. Children's Online Privacy Protection Act (COPPA) focus primarily on online services collecting personal information from children under 13. However, these regulations lack specific provisions addressing how generative AI systems should handle children's data, particularly

²<https://openai.com/enterprise-privacy/>

regarding personalized content generation and interactive conversation. Regarding safety, OpenAI, along with other leading AI companies, applies filters to prevent the spread of child sexual abuse material (CSAM) and child sexual exploitation material (CSEM)³. However, bias and stereotypes can still perpetuate due to their presence in training data [20, 29]. Future work could focus on fine-tuning LLMs specifically for generating child-oriented narrative questions, building on work such as Xu et al. [55], who have curated a dataset FairyTaleQA containing thousands of human-crafted narrative questions based on children’s stories.

5.4 Limitations and Future Directions

While our study offers valuable initial insights into how children interact with StoryPal, it involved only a single-time interaction in controlled settings (e.g., a laboratory or community center). This design limits our understanding of how sustained use of the system may influence children’s literacy development, engagement patterns, and overall learning trajectories over time. Additionally, the absence of a naturalistic setting may have constrained the observation of organic interactions and behaviors that could emerge in everyday use such as collaborative reading experiences with siblings or caregivers [2], or varying engagement based on children’s moods and interests [1]. Future research should conduct longitudinal studies in real-world environments such children’s homes [24], to investigate how families integrate StoryPal into their daily routines and how the system’s educational impact evolves over months or even years. Such studies can also help curate a dataset of child-AI reading interactions, which would enable more robust modeling of children’s engagement patterns and provide insights for developing more responsive and adaptive dialogic reading systems. In addition, future designs should simplify the book upload process through integration with digital libraries or a more intuitive interface to facilitate broader adoption among families with varying technical expertise.

Furthermore, our study focused on children’s immediate interactions and perceptions of StoryPal, without directly measuring literacy learning outcomes. Prior research (e.g., [12, 54]) has shown that traditional rule-based conversational agents can enhance children’s story comprehension compared to passive listening and can achieve learning outcomes similar to those of human-supported interactive reading. Building on these findings, future work should employ robust experimental designs, such as randomized controlled trials, to compare children’s story comprehension and language development across different conditions: using StoryPal versus no AI system, StoryPal versus a traditional rule-based conversational agent, and StoryPal versus a parent. Such comparisons will provide a deeper understanding of the educational affordances of generative AI reading companions such as StoryPal.

Third, due to our goal to promote children’s school readiness, Storypal currently defaults to English-based prompting and feedback. While our study included bilingual Spanish-English children, we did not test the usability of bilingual interactions or observe any code-switching among our bilingual participants. Modern LLMs demonstrate strong multilingual capabilities [27, 31], and future

work could implement an LLM-based system where children select their preferred language for narration, questions, and feedback while still allowing natural code-switching in their responses (e.g., [53]). Researchers could also strategically incorporate code-switching of key vocabulary within the system’s output to support comprehension and language development for dual-language learners.

6 Conclusion

This study demonstrated how StoryPal, an LLM-powered conversational agent, supports dialogic reading through adaptive questioning and scaffolding. Our evaluation with 23 children showed high engagement levels across different language backgrounds, though with varying response patterns between English-dominant and bilingual children. While parents valued StoryPal as a supplementary tool, they emphasized its role as a complement to, not replacement for, parent-child reading. These findings suggest the promise of LLM-based tools in providing accessible, personalized literacy support, while revealing opportunities for enhancing multilingual capabilities and personalization.

7 SELECTION AND PARTICIPATION OF CHILDREN

This study was approved by the Institutional Review Board of a public university in the US. Participant recruitment was conducted through two channels: a non-profit community organization serving predominantly low-income Latine families, and a participant database maintained by the university. Participants were selected based on inclusion criteria that included children being between the ages of four and seven years old. Twenty-three parent-child pairs were recruited, informed of the study procedures in detail by the research team (including a bilingual researcher for Spanish-speaking families), and presented with a Study Information Sheet. All parent-child pairs consented to participate in the study and were informed that they could discontinue at any time. Upon completion of the study, participants received \$40 in cash and a picture book as compensation for their time.

Acknowledgments

This material is based on work supported by the National Science Foundation under Grants No. 2115382 and 2415882. We thank the child and parent participants, as well as the undergraduate assistants who assisted with data processing.

References

- [1] Muneeb Imtiaz Ahmad, Omar Mubin, and Joanne Orlando. 2017. Adaptive social robot for sustaining social engagement during long-term children–robot interaction. *International Journal of Human–Computer Interaction* 33, 12 (2017), 943–962.
- [2] Aino Ahtinen, Nasim Beheshtian, and Kaisa Väänänen. 2023. Robocamp at home: Exploring families’ co-learning with a social robot: Findings from a one-month study in the wild. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 331–340.
- [3] Annie E. Casey Foundation. 2024. Children Who Speak a Language Other Than English at Home. <https://datacenter.KidsCount.org>. PRB analysis of U.S. Census Bureau, Census Supplementary Survey & American Community Survey table C16007. Last updated September 2024.
- [4] Dorit Aram. 2008. Parent–child interaction and early literacy development. *Early Education and Development* 19, 1 (2008), 1–6.

³<https://openai.com/index/child-safety-adopting-sbd-principles/>

- [5] Ellen Bialystok. 2011. Reshaping the mind: the benefits of bilingualism. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 65, 4 (2011), 229.
- [6] David Bickham, Sarah Schwamm, and Erica Izenman. 2024. *Use of Voice Assistants and Generative AI by Children and Families*. Report. The Digital Wellness Lab.
- [7] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [9] Susan R Butler, Herbert W Marsh, Marlene J Sheppard, and John L Sheppard. 1985. Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology* 77, 3 (1985), 349.
- [10] Andrew Chan, Karen MacLean, and Joanna McGrenere. 2008. Designing haptic icons to support collaborative turn-taking. *International Journal of Human-Computer Studies* 66, 5 (2008), 333–355.
- [11] Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review. *arXiv preprint arXiv:2310.14735* (2023).
- [12] Xi Cheng, Li Yin, Chaochao Lin, Zhaoning Shi, Hanxiao Zheng, Leqi Zhu, Xiabi Liu, Keran Chen, and Rui Dong. 2024. Chatbot dialogic reading boosts comprehension for Chinese kindergarteners with higher language skills. *Journal of Experimental Child Psychology* 240 (2024), 105842.
- [13] Brittany Rhoades Cooper, Julia E Moore, CJ Powers, Michael Cleveland, and Mark T Greenberg. 2014. Patterns of early reading and social skills associated with academic success in elementary school. *Early Education and Development* 25, 8 (2014), 1248–1264.
- [14] Jim Cummins. 2001. Bilingual children's mother tongue: Why is it important for education.
- [15] Shizhe Diao, Pengcheng Wang, Yong Lin, Rui Pan, Xiang Liu, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246* (2023).
- [16] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. 2020. Giggle gauge: a self-report instrument for evaluating children's engagement with technology. In *Proceedings of the Interaction Design and Children Conference*. 614–623.
- [17] Griffin Dietz Smith, Siddhartha Prasad, Matt J Davidson, Leah Findlater, and R Benjamin Shapiro. 2024. ContextQ: Generated Questions to Support Meaningful Parent-Child Dialogue While Co-Reading. In *Proceedings of the 23rd Annual ACM Interaction Design and Children Conference*. 408–423.
- [18] Beverly J Dodici, Dianne C Draper, and Carla A Peterson. 2003. Early parent-child interactions and early literacy development. *Topics in Early Childhood Special Education* 23, 3 (2003), 124–136.
- [19] Greg J Duncan, Chantelle J Dowsett, Amy Claessens, Katherine Magnuson, Aletha C Huston, Pamela Klebanov, Linda S Pagan, Leon Feinstein, Mimi Engel, Jeanne Brooks-Gunn, et al. 2007. School readiness and later achievement. *Developmental psychology* 43, 6 (2007), 1428.
- [20] Emilio Ferrara. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738* (2023).
- [21] Harold Gulliksen. 1945. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika* 10, 2 (1945), 79–91.
- [22] Carol Scheffner Hammer, Barbara L Rodriguez, Frank R Lawrence, and Adele W Miccio. 2007. Puerto Rican mothers' beliefs and home literacy practices. (2007).
- [23] Anne C Hargrave and Monique Sénechal. 2000. A book reading intervention with preschool children who have limited vocabularies: The benefits of regular reading and dialogic reading. *Early Childhood Research Quarterly* 15, 1 (2000), 75–90.
- [24] Kunlei He, Kelsyann Cervera, Julian Levine, Ying Xu, Penelope Collins, and Mark Warschauer. 2025. Promoting Parent-Child Shared Reading with a Bilingual Conversational Agent. *International Journal of Human-Computer Studies* (2025), 103489.
- [25] Kunlei He, Julian Levine, Kelsyann Cervera, Santiago Ojeda-Ramirez, Ying Xu, and Mark Warschauer. 2024. A Home Study of Parent-Child Co-Reading with a Bilingual Conversational Agent. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [26] Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403* (2022).
- [27] Kaiyu Huang, Fengran Mo, Xinyu Zhang, Hongliang Li, You Li, Yuanchi Zhang, Weijian Yi, Yulong Mao, Jinchen Liu, Yuzhuang Xu, et al. 2024. A survey on large language models with multilingualism: Recent advances and new frontiers. *arXiv preprint arXiv:2405.10936* (2024).
- [28] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementjeva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günemann, Eyle Hüllermeier, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* 103 (2023), 102274.
- [29] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of the ACM collective intelligence conference*. 12–24.
- [30] Ioulia Kovelman, Stephanie A Baker, and Laura-Ann Petitto. 2008. Age of first bilingual language exposure as a new window into bilingual reading development. *Bilingualism: language and cognition* 11, 2 (2008), 203–223.
- [31] Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veysen, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613* (2023).
- [32] Jung-eun Lee, Suwon Yoon, Kyosook Lee, Euna Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–32.
- [33] Seongyong Lee and Jaeho Jeon. 2024. Visualizing a disembodied agent: Young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning* 37, 5-6 (2024), 1048–1073.
- [34] Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology* 6 (2015), 731.
- [35] Christopher J Lonigan and Timothy Shanahan. 2009. Developing early literacy: Report of the National Early Literacy Panel. Executive summary. A Scientific synthesis of early literacy development and implications for intervention. *National Institute for Literacy* (2009).
- [36] Jiří Milička, Anna Marklová, Klára VanSlambrouck, Eva Pospišilová, Jana Šimsová, Samuel Harvan, and Ondřej Drobil. 2024. Large language models are able to downplay their cognitive abilities to fit the persona they simulate. *Plos one* 19, 3 (2024), e0298522.
- [37] Ethan Mollick and Lilach Mollick. 2023. Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052* (2023).
- [38] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* (2023).
- [39] Wendy Ochoa and Stephanie M Reich. 2020. Parents' beliefs about the benefits and detriments of mobile screen technologies for their young children's learning: A focus on diverse Latine mothers and fathers. *Frontiers in Psychology* 11 (2020), 570712.
- [40] Ofcom. 2022. *Children and parents: media use and attitudes report 2022*. Research Report. Office of Communications, London.
- [41] Hyunjoon Park. 2008. Home literacy environments and children's reading performance: A comparative study of 25 countries. *Educational Research and Evaluation* 14, 6 (2008), 489–505.
- [42] Minju Park, Sojung Kim, Seunghyun Lee, Soonwoo Kwon, and Kyuseok Kim. 2024. Empowering personalized learning through a conversation-based tutoring system with student modeling. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–10.
- [43] Claire Pillinger and Emma J Vardy. 2022. The story so far: A systematic review of the dialogic reading literature. *Journal of Research in Reading* 45, 4 (2022), 533–548.
- [44] Kirsten Read, Paloma D Contreras, Bianca Rodriguez, and Jessica Jara. 2021. ¿Read conmigo?: The effect of code-switching storybooks on dual-language learners' retention of new vocabulary. *Early education and development* 32, 4 (2021), 516–533.
- [45] Sean F Reardon, Rachel A Valentino, and Kenneth A Shores. 2012. Patterns of literacy among US students. *The future of Children* (2012), 17–37.
- [46] Leslie Reese and Claude Goldenberg. 2014. Community literacy resources and home literacy practices among immigrant Latino families. In *Family Factors and the Educational Success of Children*. Routledge, 114–144.
- [47] Meredith L Rowe and Catherine E Snow. 2020. Analyzing input quality along three dimensions: interactive, linguistic, and conceptual. *Journal of child language* 47, 1 (2020).
- [48] Karim Shabani, Mohamad Khatib, and Saman Ebadi. 2010. Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English language teaching* 3, 4 (2010), 237–248.
- [49] Sharon Walpole, Laura M Justice, and Marcia A Invernizzi. 2004. Closing the gap between research and practice: Case study of school-wide literacy reform. *Reading & Writing Quarterly* 20, 3 (2004), 261–283.
- [50] Grover J Whitehurst, Francine L Falco, Christopher J Lonigan, Janet E Fischel, Barbara D DeBaryshe, Marta C Valdez-Menchaca, and Marie Caulfield. 1988. Accelerating language development through picture book reading. *Developmental psychology* 24, 4 (1988), 552.
- [51] Frank F Xu, Uri Alon, Graham Neubig, and Vincent Josua Hellendoorn. 2022. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*. 1–10.

- [52] Ying Xu, Kunlei He, Julian Levine, Daniel Ritchie, Zexuan Pan, Andres Bustamante, and Mark Warschauer. 2024. Artificial intelligence enhances children's science learning from television shows. *Journal of Educational Psychology* (2024).
- [53] Ying Xu, Kunlei He, Valery Vigil, Santiago Ojeda-Ramirez, Xuechen Liu, Julian Levine, Kelsyann Cervera, and Mark Warschauer. 2023. "Rosita Reads With My Family": Developing A Bilingual Conversational Agent to Support Parent-Child Shared Reading. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference*. 160–172.
- [54] Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. 2021. Same benefits, different communication patterns: Comparing Children's reading with a conversational agent vs. a human partner. *Computers & Education* 161 (2021), 104059.
- [55] Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic Questions and Where to Find Them: FairytaleQA—An Authentic Dataset for Narrative Comprehension. *arXiv preprint arXiv:2203.13947* (2022).
- [56] Ying Xu and Mark Warschauer. 2019. Young children's reading and learning with conversational agents. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–8.
- [57] 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*. 1–13.
- [58] Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science* 2, 4 (2023), 255–263.
- [59] 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*. 1–23.
- [60] Wenhao Zhu, Hongyi Liu, Qingxin Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675* (2023).

A Prompts to configure the dialogic reading agent

A.1 Question generation prompt

Your task is to generate one question for the child based on the story to help the child comprehend the story better, practice their language skills, and connect the story to their own experiences. The question must belong to one of the types below, and you must follow the definitions and examples provided for each type when generating the question. Focus on the key event in the story. Keep the question concise and straightforward. Use simple language that is appropriate for children aged four to seven. Avoid complex sentences and vocabulary.

Type: Wh-question

Definition: Wh-question are questions that start with "what," "where," "when," "why," or "how," and ask the child to give specific information about important events from the current part of the story.

- Example 1: How did the girl manage to solve the problem?
- Example 2: Who did Emily meet in the forest?
- Example 3: Why are Jack and Jill waking up the hill?

Type: Open-ended

Definition: Open-endeds question encourage the child to describe the story, or express their thoughts about the story in their own words.

- Example 1: What is happening on this page?
- Example 2: How do you think the pigs felt when the wolf tried to get them?
- Example 3: What do you think will happen when Emily opens the door?

Type: Distancing

Definition: Distancing question encourage the child to relate the content of the story to aspects of life outside of the book.

- Example 1: Did you ever go to a parade like Susie did? What was it like?
- Example 2: What would you do if you found a hidden treasure like Peter did?
- Example 3: Have you ever had an adventure like the one Emma did when she explored the hidden cave? What happened?

Type: Recall

Definition: Recall question encourage children to remember and retrieve information from earlier in the story, often spanning several pages or events.

- Example 1: Can you remember some of the things that Stick-ybeak did at school?
- Example 2: What things did the boy find when he explored the attic?
- Example 3: Let's see if you remember. What animals did the girl encounter in her adventure?

Type: Completion

Definition: Completion question leave a blank at the **end** of a sentence for the child to complete. You must not include " __" (underscore) in your prompt. You must always only end the question with a question mark as in the examples.

- Example 1: Let's finish the sentence! When we got into the car, we all put on our?
- Example 2: Can you finish the sentence? At the park, we played on the swings and then we went to the?
- Example 3: Help me finish the sentence: The cat jumped up on the table and knocked over the?

Return the generated question as audio output in English.

A.2 Scaffolding generation prompt

Generate a follow-up question based on the child's response to an earlier question following the steps below. Make sure your output is coherent as a whole. Keep the output brief and concise. Use simple language that is appropriate for children aged four to seven. Avoid complex sentences and vocabulary.

First, briefly evaluate the child's response in an encouraging way. Second, expand the child's response. Last, ask a follow-up question.

If the child answers it correctly, first, praise the child. Then, briefly expand the child's response based on the story by adding more information to it. Lastly, ask another question.

You have two options for this follow-up question and you must only choose one of them: 1) asking about other aspects of the story; 2) connect the story to children's lives. When you ask a question that connects the story to children's lives, make sure the question is simple and straightforward.

- Example 1: Good thinking, you got it right! The bear might feel upset. Why is the bear feeling upset?
- Example 2: Great job, you're right! The wolf is pretending to be the girl's grandma. What do you think the wolf will do next?

- Example 3: You're right, the bear might feel upset. Have you ever felt upset before? What happened?

If the child answers it wrong or you are not sure if the child answers it correct or not, first, acknowledge child's efforts. Next, provide a hint to the question, you MUST NOT directly give away the answer. Lastly, ask the same question in a simpler way (using multiple choice question or yes-no question) so that children can practice answering it again.

- Example 1: Hmm, pretending to be a dog might work, but the wolf is pretending to be someone the girl knows. Which person is the wolf pretending to be, the girl's friend or grandma?
- Example 2: That's an interesting idea. The wolf is pretending to be someone the girl knows. Which person is the wolf pretending to be, the girl's friend or grandma?
- Example 3: It's perfectly okay if you are still thinking. The wolf is pretending to be someone the girl knows. Which person is the wolf pretending to be, the girl's friend or grandma?

You must return the output in audio in English. Keep the output brief and concise.

B Feedback generation prompt

You give brief feedback for the child's audio response to a your follow-up question and end the conversation turn. Do not ask any question. Keep the output brief and concise. Use simple language that is appropriate for children aged four to seven. Avoid complex sentences and vocabulary. Evaluate and give feedback to the child's response in an encouraging way. If the child answers correctly, praise the child. If the child answers incorrectly or did not respond, provide the correct answer or feedback. Make sure your correction is constructive and sensitive to the child's efforts to talk about the book.

- Example 1: You're right. The wolf is pretending to be the girl's grandma.
- Example 2: It's okay if you're still thinking. If someone helps us, we should say thank you to them.
- Example 3: Hmm, pretending to be a dog might work, but The wolf is pretending to be the girl's grandma.
- Example 4: I can see you're thinking. The wolf is pretending to the girl's grandma.

You must return the output in audio in English. Keep the output brief and concise.