Computers and Education: Artificial Intelligence 9 (2025) 100484

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

Leveraging an LLM-enhanced bilingual conversational agent for EFL children's dialogic reading: Insights from children, parents, and educators

Feiwen Xiao[a, b, *], Zhaohui Li[c], Jiaju Lin[d], Xiaohan Zou[e], Dandan Yang[f, g], Wenting Zou[d], Jinjun Xiong[c]

[a] Learning Sciences and Technologies Program, Graduate School of Education, University of Pennsylvania, Philadelphia, PA, United States
[b] Penn State University, Philadelphia, PA, United States
[c] National AI Institute for Exceptional Education, Institute for Artificial Intelligence and Data Science (IAD), University at Buffalo, Buffalo, NY, United States
[d] Department of Educational Psychology, Counseling, and Special Education, The Pennsylvania State University, State College, PA, United States
[e] Department of Computer Science and Engineering, The Pennsylvania State University, State College, PA, United States
[f] Center for Community Schooling, University of California, Los Angeles, CA, United States
[g] School of Education, University of California, Irvine, CA, United States

HIGHLIGHTS

- This study examines how EFL children interact with an LLM-powered bilingual conversational agent embedded in an interactive e-book, and discusses its design implications.
- The conversational agent emulates effective dialogic reading by prompting story-related questions and delivering adaptive feedback tailored to children's responses.
- Interaction with the agent enhances children's comprehension, language output, and their engagement with the story.
- The bilingual conversational agent provides adaptive and context-aware bilingual feedback that supports language learning in a multimodal digital reading platform.
- Feedback from children, parents, and educators indicates strong support for the bilingual conversational agent's educational value and suggests promising directions for future integration and research directions.

ARTICLE INFO

Keywords:
Conversational agent
Large language model (LLM)
Child-AI interaction
E-book reading
Early childhood
Multimedia learning
Interactive media
Language learning

ABSTRACT

Dialogic reading, a technique in which adults and children engage in interactive discussions around a story, has been shown to improve children's language and literacy development. Despite its evidence-based benefits, its adoption amongfamilies with English as a Foreign Language (EFL) backgrounds has been particularly challenging due to limited English proficiency, restricted conversational skills, and a low inclination to read in English. This paper presents "Storio", an e-book integrated with a bilingual large language model (LLM)-based conversational agent named "Mia", used as a design probe to investigate interactions between EFL children (N = 17) and the agent, and to gather insights from parents (N = 19) and educators (N = 2). The findings indicate that the bilingual agent effectively supports language output, fosters interactive experiences, and promotes language skills. The study offers valuable design implications for the development of LLM-based and children's interactive e-books tailored to the needs of children with diverse linguistic and cultural backgrounds.

1. Introduction

The number of children learning English worldwide has been steadily growing in recent years. English, as a *lingua franca*, has opened avenues for communication, collaboration, and access to global opportunities. As of 2020, more than 600 million people were learning English as a foreign language (EFL) (Bhattacharyya, 2021). In the United States, approximately one in five students (USA) speak English as a second language (Jiménez-Castellanos and García, 2017). Likewise, in China, the population of English learners continues to grow (He and Li, 2021). Globalization has led to growing cultural and linguistic diversity in the

* Corresponding author at: Learning Sciences and Technologies Program, Graduate School of Education, University of Pennsylvania, Philadelphia, PA, United States.
Email addresses: feiwenx@upenn.edu (F. Xiao), zli253@buffalo.edu (Z. Li), jjl7137@psu.edu (J. Lin), xfz5266@psu.edu (X. Zou), dandany1@uci.edu (D. Yang), wpz5135@psu.edu (W. Zou), jinjun@buffalo.edu (J. Xiong).

https://doi.org/10.1016/j.caeai.2025.100484
Received 6 December 2024; Received in revised form 31 August 2025; Accepted 27 September 2025
Available online 30 September 2025
2666-920X/© 2025 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

contemporary educational context, and the number of English Language Learners (ELLs) is expected to increase.

In EFL settings, however, English is rarely used for everyday communication, limiting children's exposure to the language. This lack of regular interaction with English is compounded by insufficient access to high-quality English learning resources in many contexts, including schools, afterschool programs, and community settings (Chow et al., 2008). Consequently, the language environment at home becomes a critical factor in shaping children's language acquisition outcomes (Seo, 2023). Research has identified several effective home literacy practices, including shared reading and structured language instruction, as strategies to support children's language development (Dong and Chow, 2022). Of these, story reading has proven particularly impactful, offering immersive opportunities for children to learn new vocabulary, understand sentence structures, and gain cultural knowledge (Goldenberg, 2008), while simultaneously motivating them to engage with the target language (Caesar and Nelson, 2014; DesJardin et al., 2017).

Among story reading techniques, dialogic reading has emerged as a highly effective evidence-based practice for developing children's native and foreign language skills. It is closely linked to various aspects of children's native and foreign language development, including receptive and expressive vocabulary, syllable awareness, and word reading (Chow et al., 2010; Yeung and King, 2016). It involves parents guiding their children by asking them story-related questions, giving feedback, and enriching their verbalizations (Chow et al., 2008). Despite its evidence-based benefits, dialogic reading poses unique challenges for EFL families, who often face barriers such as limited English proficiency, insufficient conversational skills, and a lack of motivation or confidence in reading in English (Chow et al., 2010).

Conversational agents (CAs), with their powerful generative capabilities and human-like conversational interfaces, have shown great potential as conversational companions to enrich the reading experience of children with EFL(Xu and Warschauer, 2020b; Xu et al., 2021a; Zhang et al., 2022). Existing research targets children's first language and literacy development, with limited attention to EFL children's learning experiences with CAs. Recent advancements in the large language model (LLM) have opened a new floor to designing human-like CAs capable of being EFL children's conversational partners during dialogic reading, which is underexplored. Moreover, in recent years, many children's language learning experiences have shifted towards interactive media, given the built-in multimedia features' affordances for second language learning.

Therefore, in this study, we designed an interactive e-book with an embedded LLM-based CA as a design probe. A user study with 17 children, complemented by in-depth interviews with the children (N = 17), their parents (N = 19) and a think-aloud study with two educators, was conducted to evaluate the usability of the design. Through interviews and observations, our study aims to uncover children's learning experiences with the CA, different stakeholders' (i.e., children, parents, and educators) insights on system design, and how future designs using LLM can be better tailored towards young ELLs.

2. Literature review

2.1. Dialogic reading for children's language development

Dialogic reading, a well-established evidence-based practice, effectively promotes language and literacy development in children by engaging them in guided conversations during storybook reading. During dialogic reading, parents typically engage children by asking them questions, giving them feedback, and elaborating on their responses (Whitehurst et al., 1988).

Two widely recognized dialogic reading techniques are CROWD and PEER (Lonigan and Whitehurst, 1998; Whitehurst et al., 1988). CROWD encourages parents to ask five types of questions: Completion, Recall, Open-ended, Wh- questions, and Distancing—to engage children in deeper thinking. PEER, on the other hand, stands for Prompt, Evaluate, Expand, and Repeat, which helps scaffold children's responses and fosters a more interactive and dynamic reading experience.

Dialogic reading has been shown to support language and literacy among native English speakers as well as language minority children developing English proficiency in English-majority settings such as the United States (Fitton et al., 2018). This conversational reading experience has shown great benefits in children's language and literacy development, including vocabulary acquisition (Hargrave and Sénéchal, 2000), oral language skills (Flynn, 2011; Kennedy and McLoughlin, 2023; Lever and Sénéchal, 2011; Mol et al., 2008) and comprehension skills (Ergül et al., 2017).

2.2. Conversational agent and children's language development

Children today frequently interact with conversational agents in their daily lives through various media, such as voice assistants (Lovato and Piper, 2015), voice-guided applications (Dietz et al., 2021), and robots (Cheng et al., 2018).

With the powerful capability of having back-and-forth conversations with humans, conversational agents have great potential to serve as conversation partners for language learning (Xiao et al., 2023). Emergent research has explored the potential of using conversational agents as children's reading conversational partners, yielding promising results. For example, Xu and colleagues (Xu et al., 2021b) investigated children's reading with a conversational agent (i.e., Google Home Mini), which asked children questions and gave them feedback during the reading session of the picture book. They found that children's reading with conversational agents could help children enhance their reading comprehension of the story. Similarly, Liu and colleagues(Liu et al., 2022) designed a conversational agent to chat with students after children finished book reading. They found that this learning experience greatly improved children's interest in reading.

Despite the benefits, most conversational agents for reading have been designed for use in children's native language. One notable exception is the work of Xu and colleagues (Xu et al., 2023), who developed a bilingual conversational agent embedded in a culturally sensitive interactive e-book to support joint parent–child reading. The agent could converse in Spanish, English, or a mix of both, and the design was shown to improve children's bilingual development and reading engagement.

Integrating conversational agents into digital learning platforms, such as e-books, offers several benefits for multilingual children. E-books are often equipped with interactive features such as narration, built-in dictionaries, and automated prompts, which enrich the reading experience and improve learning outcomes. These multimedia features provide visual and auditory stimuli that support children's multimodal sensemaking in another language (Hsieh and Huang, 2020; Xiao et al., 2023). They can sustain young learners' motivation, offer more comprehensible input, and expand EFL children's vocabulary (Yang et al., 2022). However, limited research has examined the integration of conversational agents into e-books for multilingual learners.

2.3. Leveraging LLMs to power conversational agent as children's conversation partner

Large language models (LLMs), such as GPT-4, represent a new frontier for designing more dynamic, responsive, and personalized conversational interfaces. Trained on massive datasets with billions of parameters, LLMs demonstrate strong generalization abilities and can provide personalized feedback. With simple prompt engineering, LLMs can handle many downstream tasks with a few demonstrations (Brown, 2020; Kojima et al., 2022), like sentiment classification (Miah et al., 2024), grammar correction (Fan et al., 2023), and writing improvement (Meyer et al., 2024). Unlike traditional models, LLMs-driven CAs excel in their context-aware ability. They can track the extended dialogue histories (Hsieh et al., 2024) and recognize emotional cues from speakers (Lian et al., 2024), thereby facilitating natural and dynamic conversation (Wei et al., 2022a).

Recently, both text-based and speech-based LLM-powered CAs have been implemented in educational settings. Applications include generating learning content (Dijkstra et al., 2022), task-based dialogue (Li et al., 2023), providing personalized feedback (Zou et al., 2024), and supporting language learning (Zhang and Huang, 2024). However, most of these systems targeted teenagers or adult learners. Only one study (Zhang et al., 2024) designed a CA powered by LLM to co-create Math stories to help children learn the Math language. By contrast, traditional conversational agents designed for children have been largely rule-based, relying on predefined scripts or limited sets of responses (Sabharwal and Agrawal, 2020). Such systems struggle to offer adaptive, consistent, and personalized interactions. As noted by Xu et al. (2021a), rule-based conversation often leads to pseudo-interactions, as these systems cannot adequately handle complex or unpredictable dialogue. Despite the flexibility and versatility of LLM, how they can be tailored to the conversational interface for young children from diverse linguistic and cultural backgrounds remains unexplored.

3. Method

Grounded in the well-established research evidence of dialogic reading and the powerful capability of LLM, we developed *Storio*, an e-book featuring an embodied conversational agent named "Mia". Mia was designed to be a child-friendly yellow cartoon character to engage young readers. Given the scarcity of LLM-based learning technologies for children's foreign language learning, we began the design process with a thorough review of existing literature on dialogic reading, particularly focusing on how actual dialogic reading practices support children's language learning.

Then we employed *Storio* as a **design probe** (Boehner et al., 2007) and conducted a user study involving different stakeholders and incorporating think-aloud, contextual inquiry, and interviews. This approach enabled us to explore whether Storio emulates the best practices of parent–child reading dialogic while meeting children's needs. Since this design is novel in the context of children's language learning applications, and–because few children have prior experience with such tools–we adopted a design probe methodology to enable hands-on interaction with the tool. To ensure the design probe reflects its educational value, the design process involved feedback from multiple stakeholders, including children, parents, and educators.

3.1. Designing CA-enhanced e-book as a probe

This section presents the design and implementation of our design probe, Storio, which features an e-book story with an embedded bilingual conversational agent. This e-book was designed for children learning English as a foreign language.

Design Goals (DG). The following design goals guide the creation of our design probe:

1. **DG1.** Facilitate interactive and engaging conversations that promote meaningful learning experiences and sustained attention.
2. **DG2.** Support inclusivity through linguistic diversity and multimodal communication to ensure accessibility and comprehension for all learners.
3. **DG3.** Utilize visual and interactive elements to enhance emotional support, motivation, and overall engagement in the learning process.

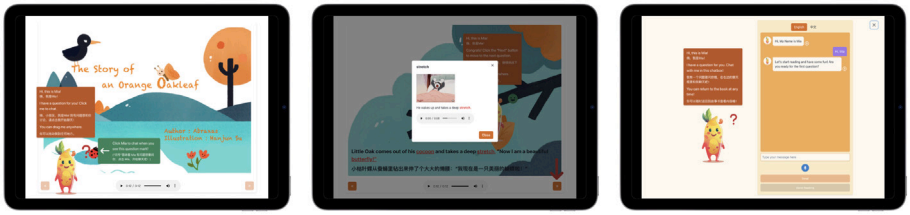
Storio as design probe. With these principles in mind, we developed *Storio*, an e-book application featuring a bilingual conversational agent, Mia, powered by a large language model (LLM). Designed for tablets equipped with built-in microphones and speakers, Mia narrates the story "Little Oak" and pauses at key moments to ask children story-related questions. Mia's conversational abilities are modeled after the feedback and guidance typically provided by parents during dialogic reading, with the added functionality of supporting bilingual interactions. To further enhance children's motivation and engagement, Mia is presented with a playful and child-friendly design (Fig. 1).

3.2. Design of LLM-powered CA

Question design. Following the CROWD dialogic reading technique (Zevenbergen and Whitehurst, 2003), we designed a total of twelve story-related questions mapping three types asked by the CA, Mia, to children during the reading session (DG1). The questions included: (1) wh-questions, including what, why, when, how (e.g., "Why did Little Oak cry again?"), (2) recall questions, which help children recall the storyline (e.g., "Why did Little Oak cry again?"), (3) a single distancing prompt question that relates children to their own experience (e.g., Have you had tears of joy?).

Scaffolding design. Two reading experts anticipated students' responses to the questions, which were then enriched with data from rapid testing sessions (N = 6). As a result, children's responses fell into the following categories: (1) correct answer and (2) incorrect answer, which also included cases when children expressed uncertainty (i.e., I don't know; I don't remember). For the first condition, Mia gives affirmation and repeats the children's responses. For the second condition, Mia offered encouragement, followed up by an explanation or rewording and a simplified question (e.g., from "why" question to a "what" question). In addition to encouraging the children to try again, Mia offered contextual clues by asking them to look at a specific place in the story. This was informed by our DG 1 and 3, which reminded children to use multimodality for meaning-making as well as maintaining their attention on the story. The final *distancing* prompt invites personal reflection ("Have you ever had tears of joy?"). When a child shares an experience, the agent responds appreciatively; if the child cannot relate, it provides a brief and supportive reassurance (e.g., "That's okay. Perhaps you will experience that emotion one day-thank you for thinking about it.").

Prompt engineering. As shown in Fig. 2, the CA was powered by a large language model (i.e., GPT-4) to support highly personalized child-agent interaction (DG1). Specifically, we applied few-shot learning (Perez et al., 2021; Wei et al., 2022b) to combine machine knowledge with human expertise. In addition to the task description, the model was fed with a few examples written by educational experts based on field tests. As in the example shown in Table 1, we ask the CA to first give corrective feedback and the corresponding scaffolding in 25 words as feedback to a child's answer to a specific question. In particular, the



The conversational agent (CA) introduces itself and prompts children to interact when needed.

A multimedia dictionary helps children understand vocabulary by providing pronunciations and pictures for key words with a simple click.

Children interact with the CA in Chinese, English, or a mix of both. Responses are then analyzed by our fine-tuned GPT-4.

Fig. 1. Screenshots of the story reading user interface of *Storio*.

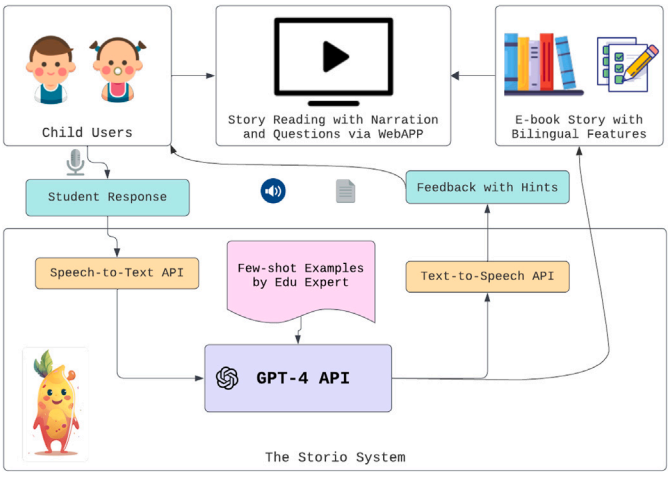


Fig. 2. The interaction flow of the Storio system.

Table 1
Chat prompt request for correctness, feedback, and hint for the answer.

Receiver	Content
Conversational agent	Question: $ {Question} Answer: ${Child's answer}. Give me correctness (Yes or No), feedback, and a hint to the correct answer within 25 words, using child-friendly language. Here are some examples: ${few-shot examples}

prompts reflected the CA's role in guiding children's responses through follow-up questions while offering constructive feedback aligned with their developmental and cognitive levels.

Design of e-book. Adhering to DG 2, we integrated Chinese translations beneath each sentence in the e-book story to facilitate bilingual understanding. To support children's sensemaking of vocabulary via multimodalities, we incorporated a multimedia dictionary for keywords, enabling children to access pronunciations and pictorial meanings with simple clicks. The CA appears within the storybook interface and can be freely repositioned on the screen. Besides, Mia is able to understand and respond in both children's native language (Chinese) and English (target language).

In line with DG 3, when it is time for interaction, a question mark flashes above the CA's head, prompting children to click on the agent to open a chat box. Children can respond in either Chinese, English or a combination of both. Audio responses are transcribed into text using OpenAI's speech-to-text API, after which our fine-tuned GPT-4 model analyzes the input and provides up to three scaffolded prompts to help children arrive at the correct answer.

After processing the input, OpenAI's text-to-speech API generates natural, human-like voice responses for children. Upon reaching the correct response, a smiling face appears beside the CA. Although Mia's default responses are in English, users have the option to access these responses in Chinese at the click of a button, powered by OpenAI's GPT-4.

3.3. User study design

To understand how Storio supports children's English reading, we conducted a user study among 17 children aged 5–8 years old. In one dyad, both parents participated with the same child, resulting in 18 parent participants in total. We also included two educators (N = 2) to provide a comprehensive view of the system's effectiveness and areas for improvement.

Participants To recruit participants, we distributed a digital flyer on social media along with our contact information. Interested parents reached out to sign up for the user study. As a result, we recruited 17 child–parent pairs to participate in our user study. Children were aged between 5 and 8 years old (Mean = 6.76, SD = 0.97). The number of females and males was evenly distributed. All the participants spoke Chinese as their first language and were learning English as their foreign language. Two educators were recruited using snowball sampling within the lead researcher's network. The detailed demographic information is in Tables 2 and 3.

Procedure All user tests included two sessions, conducted online via Zoom for participants' ease of access and convenience.

Session 1. Before each session, parent participants completed a questionnaire, reporting relevant demographic information and children's current English level (details see from Table 2). The lead researcher also provided logistical instructions for the setup. At the start of the user test, the lead researcher introduced the study procedure, and the children watched a tutorial video where Mia explained how to interact with the system. The children then read with Mia for 20 to 35 minutes, depending on their pace. Parents were permitted to be present in the room to offer technical support and observe their children's reading. After the reading session, children were asked 7 story comprehension questions about the story, followed by a 40-minute semi-structured interview with both children and parents. Chat logs from 15 children were successfully saved for conversation analysis; data from two participants were lost due to technical issues.

Session 2. Educators were first briefed on the goals of the session and asked to introduce their expertise and previous experiences, which

Table 2
Demographics of child and parent participants.

ID	Child age (years)	Parent	Parent education	Child english proficiency level	Child CA/AI exposure
P1	8	Mother	Associate	Beginner	Rarely
P2	8	Mother	Master's	Intermediate	Never
P3	7.5	Mother	Bachelor's	Beginner	Frequently
P4	7	Mother	Bachelor's	Beginner	Rarely
P5	8	Mother	Master's	Intermediate	Frequently
P6	6.5	Mother	Bachelor's	Never learned before	Rarely
P7	6	Mother	Bachelor's	Beginner	Rarely
P8	8	Mother	Bachelor's	Beginner	Never
P9	7	Mother	Bachelor's	Intermediate	Rarely
P10	6	Mother	Bachelor's	Intermediate	Rarely
P11	7	Mother	Bachelor's	Beginner	Rarely
P12	7	Mother	Master's	Beginner	Rarely
P13	5	Mother	Bachelor's	Intermediate	Rarely
P14	5	Mother	Bachelor's	Intermediate	Rarely
P15	7	Mother	Master's	Intermediate	Frequently
P16	6.5	Both	Mother: Master's; Father: Bachelor's	Intermediate	Rarely
P17	5.5	Mother	Bachelor's	Beginner	Frequently

Table 3
Demographic information of educators.

ID	Area of specialization	Highest degree
1	Bilingual Education, Language Development, Dyslexia, Instructional Design, Educational Technology	Doctor of Philosophy in Education
2	Language Development	Doctor of Philosophy in Education

helped contextualize their engagement with the system. Using the think-aloud technique, educators interacted with the system, verbalizing their thoughts and reactions in real time. The researcher occasionally asked follow-up questions to probe deeper into their experiences and observations. Each session concluded with a debriefing and a semi-structured interview to collect their impressions of the system, evaluate the current design, and identify areas for improvement. All sessions were thoroughly documented through video recordings and transcriptions, with detailed field notes taken during the tests to capture comprehensive observations and insights.

Interview protocol *Children.* We initiated the interviews by asking the children to share their overall experiences with e-book reading, their interactions with Mia, and the impact of the conversation on their learning (e.g., Does reading to Mia help you answer the comprehension questions?). Questions also explored what they enjoyed or disliked about both the e-book and Mia during their reading sessions, as well as desired additional features (e.g., "What else do you want the e-book to do?).

To gauge children's perception of the LLM-based CA, children were asked to complete the statement:"Mia is ___," allowing them to describe Mia in their own words and providing direct insights into their perceptions of the CA's personality and traits (Van Brummelen et al., 2023). Additional questions adapted from (Xu and Warschauer, 2020a) examined children's perceptions of Mia's cognitive abilities (e.g., "Do you believe Mia is intelligent?"), psychological presence (e.g., "Do you think Mia has emotions?"), and behavioral similarities to humans (e.g., "Can Mia talk?"). We also asked children to rate CA's intelligence and memory and compare theirs with the CA. Responses were typically categorized into "Yes," "No," or "I don't know," with follow-up prompts inviting children to elaborate on their reasoning. Finally, we explored children's comparative perceptions by asking them whether Mia resembles anyone they know and whether they would prefer to read with Mia or their parents in future reading sessions.

Parents. Parents were asked to evaluate their children's reading experiences with Storio, focusing on the ways the system could better support children's language development. Discussions also covered their views on the design of story-based questions, child-agent interaction, and the role of the CA, including how the CA's effectiveness could be improved.

Educators. Educators were engaged in a detailed discussion about their professional observations regarding the e-book's impact on learning. They were invited to evaluate the appropriateness of the content, the pedagogical strategies employed by Mia, and the overall educational value of the CA in the context of language learning. To elicit actionable feedback, we asked targeted questions about how the system could be refined to better align with established language education practices and curricular goals. Educators also provided feedback on how the e-book and Mia might be integrated into children's reading practices to enrich their learning experience. This feedback was used to identify concrete directions for improving the design and implementation of the e-book and CA for broader educational applications.

The present study aimed to answer the following research question:

1. How is the LLM-based bilingual CA emulating features of high-quality adult-children conversations?
2. How do children respond to the CA during the reading sessions, in terms of language production, cognitive engagement, and affect?
3. What are children's perceptions towards CA in dialogic reading?
4. How do parents, children, and educators' preferences for the conversational agent's design elements (i.e., visual, functionality, interactivity) and e-book converge and differ?

3.4. Data analysis

This section presents our data analysis methods in response to each research question.

3.4.1. RQ1. Quality of AI-child conversations

We first coded the CA-child conversation using a framework adapted from Xu and colleagues (Xu et al., 2021a) to conceptualize the characteristics of high-quality adult-children conversation. According to (Xu et al., 2021a), the high-quality adult–child conversation is typically characterized by an Initiation-Response-Feedback (IRF) cycle. Initiation refers to the adult asking children questions that prompt cognitive engagement and language production. Feedback plays a critical role in this cycle, as it involves the adult's response to the child's input, guiding and enhancing the conversation. This cycle encompasses several distinct feedback types, which are elaborated below. The coding framework, along with corresponding illustrative examples, is presented in Table A6.

Initiation - initiation refers to the opening move of an interactional sequence, typically realized through the CA's first question in each conversational turn. In our study, initiation is not coded because the first question for each conversational turn is preset. However, these initiations were still marked in the transcripts to provide context for subsequent analyses.

Feedback - Feedback was analyzed across several dimensions, including contingency, extension, encouragement, and adjustment.

Contingency was coded when the CA's feedback was tailored to the child's specific response. Contingency was coded as present if coders determined that the feedback specifically addressed the child's response logically instead of merely providing generic or unrelated feedback that was not based on the child's response.

Extension captured how the CA elaborated on the child's input to deepen the conversation; Extension was coded as present if the feedback contained additional information instead of merely verifying the child's response.

Encouragement refers to how the CA motivated the child to engage further in the conversation, and includes two circumstances. It was coded as present if the feedback followed a child failure to answer the question and encouraged the child to respond to the prompt; when the child's response was not correctly identified as correct forms by the voice interface, the agent repeated or rephrased the child's response with correct grammar, giving the child an example to follow without overt criticism.

Adjustment refers to the ways in which the CA tailored its feedback to align with the child's developmental level and current proficiency level. It included five circumstances: *(1) explicit correction*. This involved the CA explicitly identifying and correcting mistakes in grammar, vocabulary, or meaning to guide the child toward a more accurate response, *(2) simplification of the questions*. This occurred when the CA simplified complex questions or concepts by providing more accessible explanations, or breaking them down into smaller and more manageable pieces; adjusting the difficulty level of questions based on the child's responses (e.g., change from Why-question to What-question), *(3) contextual cues*. This was coded when the agent referred to illustrations or the context of the story to help guide the child's thinking and to remind the child to refer back to the previous storyline (e.g., "Look at the picture in the story: What do you see on the wings of the butterfly?'). *(4) bilingual engagement* refers to the agent promoting language switching; when the children didn't speak English, the CA encouraged children to speak English; the CA explicitly showed translanguaging practices to children. This code captures the CA's use of multiple languages within the same instructional interaction to aid comprehension, reinforce learning, and connect

linguistic knowledge across languages. (e.g., Good job! The blackbird was sleeping when Little Oak arrived. ``午休'' is Chinese for "taking a nap"). Examples for each feedback dimension can be found in Table A6.

Children's input Moreover, children's responses were qualitatively coded into three categories: correct, incorrect and unclear. Correct indicated that children answered the CA's question accurately and relevantly. Incorrect responses included answers that were factually wrong, off-topic, or expressions of uncertainty. Partially correct responses were those that contained elements of correctness but were unfinished or grammatically incorrect, thus affecting the overall meaning. Unclear refers to responses that the CA failed to identify correctly.

The unit of analysis is each instance of feedback provided by the CA and each utterance by the child. Two coders discussed the coding framework and refined it iteratively during several rounds of discussion, and separately coded 20 % of the data (Cohen's Kappa = 0.86). Then, they discussed discrepancies in coding and coded the rest of the data. Once finished, they cross-checked each other's coding and discussed disagreements until an agreement was 100 % reached. Drawing on Heritage's conversation analysis (2004), we calculated the occurrence of each coding category and examined the frequency of question-answer adjacency paired actions, which typically follow a sequence pattern that structures interaction. These patterns often feature a predictable organization that drives task completion and illuminates the progression of dialogue. To enhance the interpretive depth of the analysis, field notes were incorporated, enabling triangulation across quantitative coding, qualitative interpretation, and contextual evidence. Note that the CA's final distancing question didn't enter this part of analysis.

3.4.2. RQ2. Children's responses to the conversational agent

Children's dialogic reading sessions were video-recorded and transcribed to examine their interactional patterns with the CA. Inspired by the prior research(Heritage, 2004; Kasper and Wagner, 2014; Nakamura et al., 2021; Philp and Duchesne, 2016; Ruusuvuori, 2012; Xu and Warschauer, 2020a), the analysis involved children's both verbal and non-verbal responses to the CA. Three coding dimensions were included: language production, cognitive engagement, and affect. Together, these dimensions provided a comprehensive account of how children engaged with the CA in a language learning context, encompassing their communicative behaviors, attentional involvement, and emotional engagement.

Language production The first coding dimension captures children's verbal responses to CA, which constitutes a prerequisite for sustaining conversation (Xu and Warschauer, 2020a). Three metrics were included to represent conversation production: (1) the response rate for questions. This refers to the proportion of questions posed by the conversational agent (CA) to which the child provided an answer; (2) the number of utterances. An utterance was defined as a unit of spoken language produced either by the child or the CA; (3) the word count per utterance; and (4) the number of turns of conversation. A conversation turn refers to a complete exchange between the child and the CA, where one participant initiated, and the other responded. Each turn marks a shift in the speaking role, capturing the back-and-forth interaction between the child and the CA during the dialogic reading session. Finally, to account for bilingual use, each utterance was classified into three categories: Chinese, English, and Mix of Both.

Cognitive engagement Cognitive engagement refers to the mental effort children invest in reading. In line with prior work(Nakamura et al., 2021; Philp and Duchesne, 2016), this dimension explores to what extent children interact actively and thoughtfully with the CA's prompts. We coded instances where children indicated problem-solving, sustained attention, and the depth of their thinking when responding to the CA. Problem-solving refers to children's attempts to overcome communication difficulties; Sustained attention is demonstrated through children's consistent focus on the storyline; Depth of thinking is evident when children provide detailed explanations or make inferences during their dialogic interaction with the CA. In addition to these behavioral indicators, children's English reading comprehension was assessed immediately after the user study and incorporated as a further indicator of cognitive engagement.

Affect The affect dimension refers to children's emotional responses during interactions with CA. Prior studies on child-agent interaction have emphasized this dimension in unpacking children's learning and interaction. In line with existing literature (Kasper and Wagner, 2014; Xu and Warschauer, 2020a), a speaker's affect was operationalized through multimodal cues, including facial expressions, tone of voice, and body language, which were used to determine how positively or negatively children responded to the CA's questions and feedback. Observable emotional cues such as smiles, laughter, frustration, or disengagement were used to assess the children's emotional involvement in the interaction. Each instance of affective response was coded, supplemented by detailed observational notes to provide additional contextual interpretation.

Beyond the coding of each dimension, an additional layer of observational analysis was conducted, drawing on the principles of conversation analysis (Heritage, 2004; Kasper and Wagner, 2014). This analysis examined how verbal and nonverbal cues, including gaze, gestures, body movements, and facial expressions, intertwined within each instance of the above three categories. It also considered broader contextual factors, including local and global sociocultural influences, that shaped children's responses.

3.4.3. RQ3. Children's perceptions of the CA

Children's perceptions of the CA were examined across cognitive, psychological, and behavioral dimensions. Descriptive statistics were used to analyze this data. To gain deeper insights into children's engagement and reactions, interviews about their overall experience with the Storio were transcribed, coded, and thematically analyzed. In addition, field notes, children's drawings, and observational data were integrated to provide a comprehensive understanding of how children interact with the CA and the system during reading sessions. This triangulation of data sources enriches our analysis of the learning outcomes associated with the use of Storio.

3.4.4. RQ4. Different stakeholders' perceptions of storio

This research question examined how different stakeholder groups perceived the use of Storio. Semi-structured interviews were conducted with participants from each group and subsequently transcribed and coded. The analysis focused on identifying points of convergence and divergence in their perspectives, with particular attention to how each group valued Storio's role in supporting children's reading experiences and learning outcomes.

4. Findings

4.1. How does LLM-based CA emulate high-quality adult–children conversations?

4.1.1. General distribution of children's responses and CA feedback

As shown in Table 4, approximately half of children's responses (50.6 %) are correct, indicating that story difficulty was appropriately calibrated. The remaining responses are categorized as 'Incorrect', 'Partially Correct', and 'Unclear', where 'Unclear' accounts for almost half of them.

As for CA's feedback, almost all of the feedback was contingent upon children's responses (98.86 %), indicating strong adaptivity in generating contextually appropriate responses and sustaining a dynamic conversation flow. Among specific feedback strategies. *Encouragement* was the second most frequent type of feedback at 25.76 %, followed by *simplification* and *contextual cues* at 22.35 % and 17.05 %, respectively. Overall, different feedback types were evenly distributed. Notably, despite not explicitly incorporating bilingual encouragement examples into the prompt, the CA unexpectedly employed bilingual cues in 10.98 % of

Table 4
Children's responses and CA's feedback.

Coding item	Count	Percentage
Child Response		
Correct	132	50.60 %
Incorrect	40	15.30 %
Partial Correct	30	11.50 %
Unclear	59	22.60 %
Total	261	
AI Feedback		
Contingency	261	98.86 %
Extension	53	20.08 %
Encouragement	68	25.76 %
Explicit Correction	47	17.80 %
Question Simplification	59	22.35 %
Contextual Cues	45	17.05 %
Bilingual Engagement	29	10.98 %
Total	264	

cases to enhance children's sensemaking of the story, either by encouraging English use or integrating children's home language to support comprehensible input.

Taken together, these patterns suggest that the CA flexibly integrates linguistic and interactional strategies to scaffold child-agent dialogue during dialogic reading.

4.1.2. Interaction patterns between children and CA

We analyzed the conversational patterns between children and the CA by cross-tabulating the CA's feedback types against the children's answer types. The results are visualized as a heatmap in Fig. 3. Each cell shows the percentage of times a specific CA feedback type (T_{CA}) was given in response to a particular child answer type (T_{CHI}). This percentage is the count of that specific feedback-answer pair, normalized by the total count of that child's answer type.

Our analysis reveals distinct clusters of feedback types associated with different types of child responses. Following correct responses, the CA predominantly provides feedback characterized by high *contingency*(98.44 %), *extension*(37.01 %), and *bilingual engagement*(13.39 %). This pattern suggests that once the children provide correct answers, the agent is able to build on the child's contribution to advance the children's current knowledge. This scaffolding pattern is unexpectedly consistent with Krashen's input hypothesis, which posits that language acquisition occurs when learners are exposed to language slightly above their current level. In contrast, incorrect or unclear responses trigger a broader and more varied set of feedback strategies, including *encouragement*, *explicit correction*, *question simplification*, and the provision of contextual cues, aimed at repairing understanding and supporting task completion.

We also compared interaction patterns by English proficiency (Figs. 4 and 5), grouping children using parent-reported levels: beginners or lower formed the low-proficiency group (N = 9), and intermediate or higher formed the high-proficiency group (N = 8). Three differences emerged. First, CA's *bilingual engagement* exclusively triggered in lower-proficiency groups (17 out of 17 cases). This pattern is consistent with more frequent use of Chinese in this subgroup and the CA's tendency to prompt English use or supply comprehensible input via the home language. Even if they got correct answers, the CA tended to encourage them to speak English. Second, the share of extension for the high proficiency group (47.69 %) was almost twice that of the low proficiency group (25.00 %), suggesting that richer and more accurate child productions afforded the CA more opportunities to elaborate. Third, in response to children's unclear responses, children with lower proficiency received more question simplification, while high-proficiency children received more encouragement. For low-proficiency learners, the CA maintained simpler or shorter exchanges to avoid overwhelming them.These findings indicate that the CA adapted its scaffolding strategies to learners' proficiency, balancing repair and advancement to sustain productive dialogue.

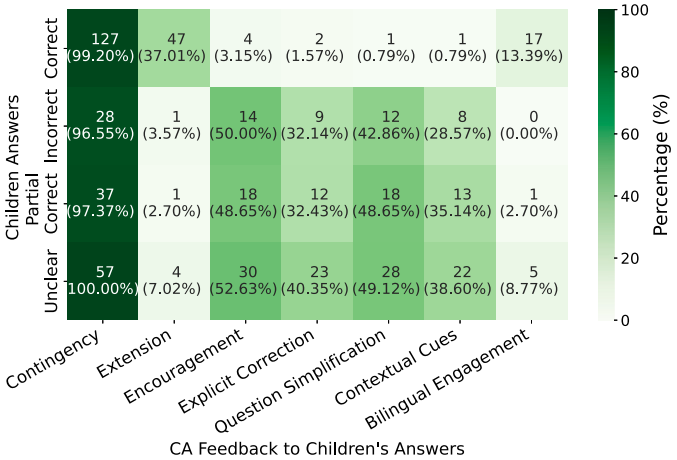


Fig. 3. General CA feedback patterns to children's responses.

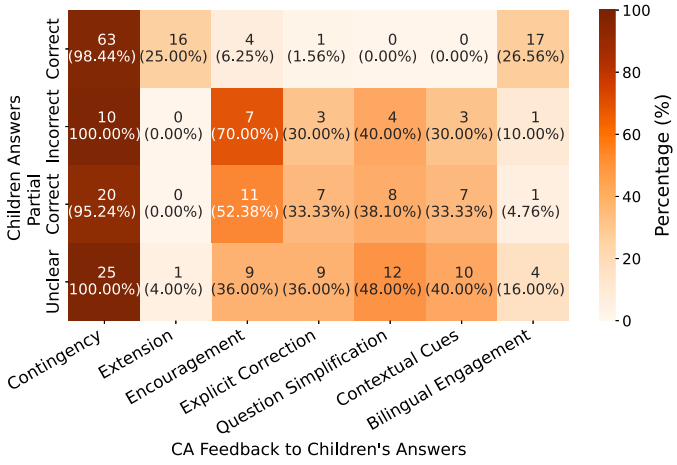


Fig. 4. CA feedback pattern to children of lower english proficiency.

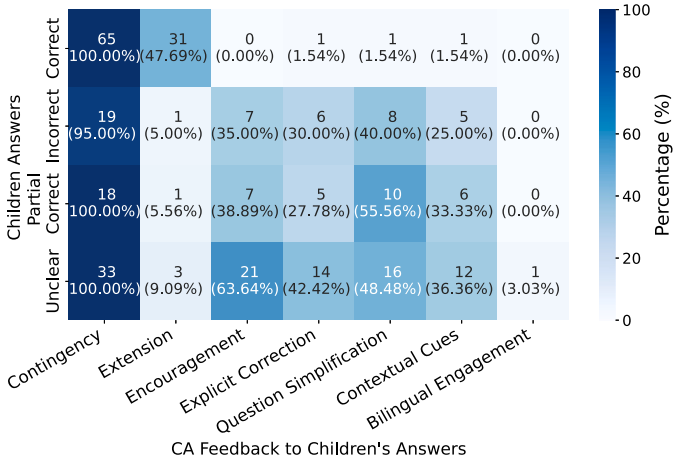


Fig. 5. CA feedback pattern to children of higher english proficiency.



Fig. 6. Concurrency of CA feedback.

4.1.3. Concurrent feedback type

To explore what kind of feedback type tended to occur within the same CA turn, we computed pairwise concurrency among feedback categories and visualized the results in Fig. 6. Each cell reports the frequency with which two feedback types appear simultaneously in a single CA response.

A high concurrency of contingency with all other categories indicated a highly dynamic and flexible conversation flow. Besides, we observed a high concurrency rate of three of four *adjustment* types, *explicit correction*, *question simplification*, and *contextual cues* share high concurrency, indicating a strong likelihood of any two of them being employed in tandem. However, *bilingual encouragement* shares little concurrency with other categories, except for *contingency*. This observation aligns with earlier findings indicating that 'bilingual engagement' is more likely to be used to provide feedback on children's correct responses. This aligns with earlier findings that it is more likely to follow correct responses, for example, when a child answers accurately in Chinese, the CA often encourages subsequent responses in English. Similarly, *extension*, which mirrors the characteristics of *bilingual engagement*, predominantly appears in scenarios where children provide correct answers.

According to the comprehensive analysis of the interaction data, our LLM CA adeptly employs a variety of feedback strategies tailored to the children's proficiency levels and response types, ensuring that interactions are not only relevant but also pedagogically supportive. The high rates of contingency feedback show the CA's ability to provide pertinent responses, thereby facilitating an adaptive conversational flow comparable to that observed in high-quality adult–child interactions. Importantly, the CA differentiated its feedback strategies based on different types of children's responses: less accurate answers typically prompt explicit correction and contextual cues, while accurate responses elicit extensions and bilingual encouragement. Such responsiveness reveals the CA's capacity to approximate the interactive qualities of skilled human conversational partners, offering scalable yet personalized support. These findings suggest that LLM-based CAs have great potential to model the nuanced, responsive, and developmentally appropriate exchanges that are characteristic of effective pedagogical dialogue.

4.2. How do children interact with storio?

4.2.1. Language production and bilingual use

In this study, children (N = 15) responded to all the questions posed by the agent during the reading sessions, producing 289 utterances across 650 turns of child-agent conversation. Children spoke Chinese, English, and a mix of both languages, yielding a total of 890 Chinese characters and 1101 English words.

More than half (53 %) of children engaged with the CA only in English,10 % communicated exclusively in Chinese, and 26.7 % used both languages.Children,who used only English, produced, on average, 91.75 English words and 7.3 words per utterance. Those who used only Chinese produced 137 characters per session and 8.3 words per utterance. This suggests not only a tendency toward longer utterances but also greater comfort or fluency when expressing ideas in Chinese. Children, who spoke in a mix of both languages, used fewer English words (4 per utterance) and more Chinese characters (10 per utterance).

Children whose parents reported early English learning experience were more likely to speak English. Those who didn't receive substantial early training had difficulty producing complete English sentences, and they tended to speak a mix of Chinese and English. Although participants were informed that code-switching was acceptable, only one child (P11) used code-switching strategies to respond to the agent (Table A7).

This limited incidence of code-switching contrasts with the findings from another study examining children's interactions with Spanish-English conversational agents (Xu et al., 2023). Two factors may help explain the difference in our setting. In our study, the CA posed questions solely in English or Chinese, despite offering bilingual options. The CA occasionally encouraged the child to respond in English, which potentially shaped children's language choice. Second, children appeared to develop a mental model of interacting with the CA from the agent's turn-taking and feedback patterns (Xu and Warschauer, 2020a), and from the sociocultural norms surrounding language use in the EFL contexts, where English is often preferred in educational or formal settings (Jin, 2005). Consequently, children are likely to adhere to these norms and continue these practices when interacting with an English-speaking CA, exhibiting language choices that paralleled expectations in formal instructional settings.

4.2.2. Cognitive engagement

During the reading sessions, most children demonstrated robust cognitive engagement, evidenced by several distinct behaviors. They sustained visual attention on the e-book, managed device navigation independently, and engaged deeply with the content. Off-task behaviors were rare, aside from occasional background noise that sometimes interfered with speech recognition. By design, the e-book requires the children to control navigation, affording them agency over their learning process. Notably, in several cases (e.g., P11, P12), children actively resisted parental intervention—physically moving a parent's hand aside—to manage the task autonomously. Such actions reflect self-directed engagement and task ownership, rather than passive or surface-level participation.

Most children also actively used problem-solving strategies to repair conversational breakdowns with the CA. Following recognition or comprehension errors, many proactively navigated back to relevant pages or prompts to re-engage with the task. More proficient children (e.g., P2, P3, P9, P13, P14, P17) frequently re-listened to narration to extract information, relying on their auditory processing skills to verify their understanding. In contrast, less proficient children (e.g., P11, P12, P16) drew on Chinese translations to scaffold comprehension before returning to the English text. This highlights a strategic use of bilingual resources to scaffold their meaning-making.

Children also employed metacognitive regulation strategies that not only facilitated immediate task completion but also contributed to longer-term learning processes. Advanced learners (e.g., P9, P10) adapted their ways of communication based on their mental model formed from prior interactions with the CA, using strategies like rehearsing answers and creatively applying new vocabulary in new contexts (detailed examples see Table A7). Less advanced learners (P11) relied on repetition and in-situ CA feedback to practice pronunciation and vocabulary (detailed examples see Table A7).

Overall, these behaviors indicate that children are not passive recipients of information. Instead, they actively monitor their learning and performance, adjust strategies in real time, and leverage technological and linguistic tools to sustain mutual understanding and progress. This pattern underscores the design's potential to elicit and support high levels of cognitive engagement during child-CA dialogic reading.

4.2.3. Affect when responding to the agent

During interactions with the CA, children displayed a range of emotional reactions, highlighting the importance of designs that are both emotionally and culturally responsive.

Across sessions, most children exhibited neutral affect, consistent with prior work findings from another study focused on younger children's literacy interaction with conversational agents (Xu and Warschauer, 2020a). This neutrality likely reflects the low-drama nature of the narrative and the CA's prompts primarily story-based prompts, both of which are less likely to elicit strong emotional responses.

Affective displays varied by age groups. Younger children, such as P7, more expressively and openly demonstrated positive emotions, including laughter and playful singing, whereas older children showed more complex emotions shaped by the CA's accuracy. For example, one child responded with restrained laughter to a misrecognition error, while another (P4) expressed frustration by pounding the table, which was alleviated by his mother's reassurance. Additionally, P9's covered her eyes while waiting for the CA's responses. This indicated a mix of anticipation and anxiety, possibly due to previous experiences with slow or inaccurate responses.

These patterns should be interpreted within the sociocultural context of learning English as a Foreign Language (EFL) among Chinese children, where English usually functions as an examination subject and carries heightened evaluative pressure. Despite the CA's encouraging scaffolds, children's affective responses sometimes reflected the exam-oriented pressure and exam-oriented anxiety associated with English learning (MacIntyre and Gardner, 1994). This highlights the importance of designing CAs that are not only technically competent but also emotionally and culturally responsive to the learning context.

4.3. What are children's perceptions towards CA in dialogic reading?

4.3.1. Cognitive, psychological, and behavioral dimension.

As shown in Table 5, most children perceived Mia, the CA, as intelligent, with 88.2 % affirming this characteristic, whereas 12.8 % did not. When asked, "Who is smarter? You or Mia?", about half of the children (41.1 %) indicated that Mia was smarter than them,three of them were unsure,and an equal share (41.1 %) judged themselves smarter. The primary justification for considering Mia smarter was that "Mia can speak English better than them" (P7). Children who believed Mia was alive or "semi-alive" (N = 5) also tended to rate her as more intelligent. Conversely, children (41.1 %) who believed they were more intelligent than Mia often critiqued Mia's capabilities, stating, "Mia is intelligent but not always". This perception, particularly from younger children (aged 6) like P6, who had never learned English before, and whose Chinese was often recognized incorrectly by Mia. One children (P1) characterized Mia as "AI", noting that "Humans are better at recognizing speech accurately. AI will make mistakes." Despite this, this child also believed that Mia was more intelligent than him, citing Mia's language ability and efficiency in providing detailed responses. Overall, perceptions of Mia's intelligence varied with children's individual experiences of the CA's language capabilities. This is in line with prior work that shows children use language abilities as a heuristic for gauging cognitive capabilities (Xu and Warschauer, 2020b). This is further supported by children's perception of Mia's behavioral capability to talk (94.1 %) and to listen (82.4 %).

Table 5
Children's perspectives on CA from cognitive, psychological, and behavioral dimensions.

Dimension	Aspect	Yes	No	I Don't know
Cognitive				
	Intelligence	15 (88.2 %)	2 (12.8 %)	0
	Memory	10 (58.8 %)	2 (11.8 %)	5 (29.4 %)
Psychological				
	Emote	4 (23.5 %)	9 (53 %)	4 (23.5 %)
Behavioral				
	See	7 (41.2 %)	7 (41.2 %)	3 (17.6 %)
	Listen	14 (82.4 %)	3 (17.6 %)	0
	Talk	16 (94.1 %)	1 (5.9 %)	0

Perception of Mia's memory were mixed, with 58.8 % affirming Mia's capability to remember. This variation in perception likely reflected children's observations of whether Mia referenced earlier turns or carried context across prompts. Although the CA sometimes produced contextual cues during dialogic reading, this behavior was not universal to children of different English levels. As a result, judgments often hinged on whether Mia appeared to recall previous inputs. Because the ability to communicate effectively is a salient marker of intelligence for children, they extended this criterion to the CA. For instance, one child judged Mia's memory was better than herself because Mia distinctly remembered the colors of the butterflies but the child could not. Taken together, these accounts suggest that children evaluate intelligence based on the consistency of Mia's responses and her ability to recall context from previous interactions, further highlighting the importance of developing CAs that can manage contextual memory to foster smoother and more coherent interactions.

Moreover, although only a small proportion of children explicitly affirmed (23.5 %) that Mia had emotions, other children (P5) described scenarios where Mia displayed happy or questioning expressions in response to their responses, and ascribed "mean" and "shy" to Mia (P6). Another kid (P10) justified that because Mia could ask questions, she must therefore have emotions. This suggests that children anthropomorphized Mia, treating her not merely as a functional tool but as a social entity whose behaviors carry symbolic and relational meaning. The findings suggest a need for CAs that can exhibit more expressive and contextually appropriate emotional behaviors.

4.3.2. Aspirational CA and e-book design

Interviews with children surfaced several themes regarding Mia (the conversational agent) as a reading companion (Fig. 8). Most children (75 %) likened Mia to a familiar figure—either their mother or a teacher—citing similarities in how these figures guide their reading by asking questions and providing assistance. One child joked that although Mia behaved like a teacher, she "looked like an ice cream." When asked to choose between reading with Mia or with their parents, a substantial majority (71.4 %) preferred Mia, often because she could communicate in English when their parents could not and because interacting with Mia felt "more fun" (e.g., P16). The remaining children reported no strong preference, noting that "both can correct me when I make mistakes" (P8).

Children alsoreported conversational difficulty due to accented speech and speech recognition errors. One child reported that "when I said Little Oak, Mia took it as Little Ugly" (P9). Another child said, "Mia cannot hear me clearly. When I said something, she heard something else" (P10). Despite such issues, most children expressed interest in reading with Mia again. A Child (P17) even drew out a podium ranking her favorite e-book so far, where Storio ranked right after Khan Academy (Fig. 7, Number 4). Only two children declined, with one (P13) stating a preference to play with their Baby Spiderman toy instead, illustrating a typical developmental inclination towards familiar and tangible playmates (Druga et al., 2018). Further, after engaging in story-based questions with Mia, over 90 % of the children correctly answered all the comprehension questions posed by researchers, indicating that reading



Fig. 7. Children's drawings of aspirational CA.

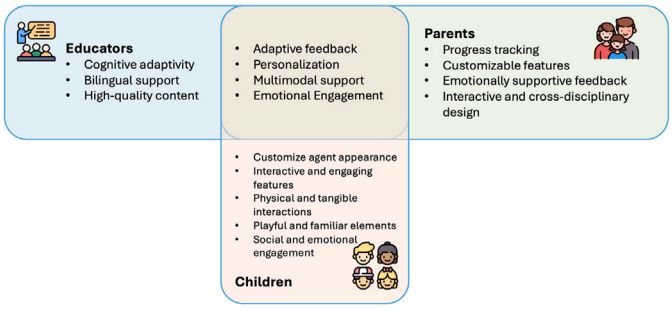


Fig. 8. Different stakeholders' design preferences.

with Mia not only enhanced their comprehension but also their ability to recall details of the story.

Additionally, when invited to redesign Mia, children articulated and sketched their ideal conversational reading partner. Themes from these redesigns emphasized richer interaction modalities, enhanced personalization, and features that would foster a greater sense of agency and entertainment during reading sessions.

Personalization. Children demonstrated a strong interest in customizing Mia's appearance, suggesting modifications that would make Mia more visually appealing and relatable. Several boys expressed a preference for transforming Mia into a robot (P17, P6, P15), citing robots' sounds as more humorous and visually appealing. P6 passionately proposed the look of an astronaut, as it was relevant to how his name sounded, "I'd prefer it to be designed as a handsome astronaut. I really think so and my 'name' stands for aerospace, so I like it." Many girls advocated for Mia to adopt a more colorful appearance, especially favoring pink and other bright colors. They envisioned Mia "with pink hands that could change colors" (P7), adding a magical touch to the interaction. Some girls mentioned a colorful cat (P7) and "yellow duckling" (P16), reflecting preferences for cute, friendly, and emotionally resonant characters. Several other girls envisioned Mia as her friend of their age (P11, P9, P13). These suggestions illustrate children's desire for personalization not only to enhance visual appeal but also to foster emotional connection and identity reflection.

Agency and interaction modality. During our study, children demonstrated a strong desire for greater agency in their interactions with Mia, the CA, and the stories she narrates. This was particularly evident in their requests for more diverse ways to interact with the narrative and characters. For instance, children (P8, P16) expressed a wish to "draw the rainbow" within the narrative. This signaled a need for more interactive and flexible input modalities beyond the primary mode of verbal communication. Additionally, one child's expressed desire to "friend the Little Oak (the story's main character) and Mia" (P8), framing the CA as a potential social companion rather than a purely computational entity. This points to a conceptual shift wherein children perceive CAs as relational entities with whom they can form social connections. Designing future systems that offer customizable, creative, and socially responsive interaction pathways—potentially incorporating affective computing, multimodal input, and co-constructive narrative elements—could thus enhance engagement and support more meaningful learning experiences.

Embodiment and tangibility. Children expressed a clear preference for tangible, physically manipulable forms of the CA, underscoring the importance of embodiment in learning and play. For instance, P6 wished to physically pull Mia out of the digital screen, suggesting a yearning for more tactile and tangible engagement with the CA. This desire for tangibility was further echoed by another child who envisioned Mia in the form of beloved furry toys like Pikachu, a unicorn, or Spiderman. Such preferences highlight children's natural inclination towards tangible, touchable entities that they can interact with physically, akin to their interactions with traditional toys. Such preferences point to broader potential of embodying digital entities embodying digital entities, such as LLM-based CAs, in familiar and physically interactive forms. Integrating elements that bridge the digitalnarratives and real-world interactions can make children's learning experiences more engaging, relatable, and effective.

Enhancements. Children proposed several functional improvements. P1 articulated a desire for Mia to provide more structured vocabulary support. Specifically, he suggested that Mia could introduce and explain difficult vocabulary at the beginning of each story session. Although Mia employs a natural text-to-speech voice, multiple children perceived it as monotonous and insufficiently expressive. Two kids (P17, P11) mentioned that a more "robotic" or characterful voice might be more entertaining, indicating a preference for greater prosodic variation, distinctive character voices, and adaptive expressivity to sustain engagement.

4.4. How do different stakeholders' preferences for design elements converge and differ?

4.4.1. Overall feedback from parents

First of all, based on parents' feedback, we found the current system largely met their expectations of CA-enhanced dialogic reading and most perceived clear benefits for their children. They affirmed the value of CA in helping children with both English reading comprehension and oral languagedevelopment (P7, P10), and described it as serving the CA by providing humanlike, parent-like guidance through the reading experience. As P4 mentioned, "I am in an assisting role. If AI in the future is accurate enough, gradually I won't be needed." Also, the guidance the CA provided to children was "interactive" (P8) and guided children to "deeply reflect on their learning" (P15). Another parent (P9) also mentioned that Mia's "feedback was quick and efficient, and more capable than theirs, as "it's best just to let my kid converse in pure English, I think this would be better, as we aren't really able to converse with her." P16 mentioned that "compared with practicing English with native speakers, in terms of both time and cost, this application is definitely a more universal and affordable option."

4.4.2. Visual

Although children showed different preferences for what Mia should look like, they all mentioned something relatable to themselves, either hobbies, favorite animals, or cartoon characters. Parents, though mostly satisfied with the current visual design, saw the potential of personalizing the CA's appearance to boost motivation and engagement. For example, P7 also mentioned, "Children love doing dress-up games,

so it would be a good idea to allow children to dress up Mia… probably rewarding children with basic accessories associated with their learning performance." This approach will gamify the experience as well as make the reading experience more personalized. However, an educator (E1) focused on the usability and cognitive implications of the CA's visual design. The educator suggested that the CA should integrate more seamlessly into the reading environment, acting as an unobtrusive guide rather than a disjointed character, to reduce cognitive load and enhance the natural flow of interaction. Aligning the CA's appearance with the characters in the story could make the learning experience more intuitive and immersive, avoiding the sense of disconnection that can occur with more abrupt interactive designs.

4.4.3. Functionality

Parents and educators primarily focused on effective feedback mechanisms to directly or indirectly facilitate children's language learning. They expressed a desire for features that could provide both direct feedback to children and progress tracking for parents. For instance, P8 noted the importance of monitoring children's engagement, stating, "If he's holding a tablet, he might get distracted by other things. You wouldn't know if he's finished reading. Is there a reading report to track his progress?" This echoed educators' (E1, E2) suggestion for an assessment feature that offers real-time feedback on the pronunciation of every single word and sentence in the narrative. Moreover, feedback tailored for younger children was suggested to be more engaging and diverse, a parent (P8) also emphasized that feedback for younger children should be diverse, utilizing sounds and animations rather than just verbal responses. A child suggested interactive features such as "eyes that change color based on whether an answer is correct (blue) or incorrect (red)" (P15). While children also wanted more features, their requests were essentially focused on enhancing their interaction with digital content. At the emotional level, P8 noted that corrective feedback can sometimes feel harsh. She therefore recommended that CA is expected to interpret the children's non-verbal cues—such as facial expressions or vocal tone—before delivering negative feedback, so that mistakes are acknowledged without causing embarrassment or discouragement. Many parents also expressed the desire for CAs to use emotionally expressive tones, making the interaction more engaging and enjoyable for children.

Bilingual support in e-books, such as multimedia dictionaries and Chinese translations, was acknowledged by most parents and all educators as beneficial, though some parents expressed concerns about potential overreliance on native language. Despite these reservations, the overall consensus favored bilingual features as flexible tools that, when used judiciously, promote deeper engagement and story understanding. To address this balance, Educator (E2) suggested a implementing a feature that allows caregivers or educators to adjust the proportion or blend of English and Chinese within the story content. Such configurability would enable differentiated support based on learners' proficiency levels and promote gradual transitions toward greater target language exposure. In addition, pronunciation assistance through bilingual scaffolding was noted as especially helpful for supporting children's oral language development.

Both parents and educators mentioned ways to personalize the current design to support children across diverse language proficiencies. Suggestions included adding an optional functionality layer, for example, incorporating leveled reading features based on reading assessment scales (P13, P14, P16, P10), multiple-choice formats to mitigate speech recognition challenges (P9, E1, E2), given that current language models sometimes struggled to accurately recognize children's speech sometimes. These adjustments aim to enhance accessibility and reduce potential frustration during interaction with the CA.

Parents and educators also emphasized the importance of creating designs that are cognitively and developmentally appropriate, as well as culturally and emotionally responsive to their children. P13 advocated for integrating Chinese cultural stories and scientific education into digital narratives, noting that using English to tell these stories can foster both cultural connection and global awareness. P7 highlighted the need for content that aligns with children's developmental stages and suggested a multidisciplinary approach that could be beneficial for early learners. As for cognitive levels, educators (E1, E2) emphasized that illustrations should closely match the content to minimize cognitive load and improve comprehension. Additionally, educators (E1, E2) and P7, a parent as well as a former teacher, emphasized the importance of repetition in learning, noting that children benefit from moving gradually from single words to fluent sentences.

Lastly, both children and parents emphasized the need to bridge the digital experience with real-world interaction, extending learning beyond the screen, though their motivations differed. Children expressed a desire for tangible and interactive elements that would allow them to engage physically with the content. For instance, they wanted to "draw a rainbow" or "customize Mia" within the narrative (P7, P16), seeking hands-on activities that would make the digital reading experience feel more like playing with traditional toys. A parent (P15) supported this, suggesting that CAs could prompt children to draw at key moments in the story, and children could change how the story proceeds through drawing.

4.4.4. Interactivity

Parents and children alike expressed a desire beyond simple question-answer exchanges. They envisioned richer and dynamic experiences involving animations and real-time visual responses that evolve with the conversation. Children, in particular, want tangible elements they can engage with and build connections to, fostering a sense of socialization within the digital environment. Some parents, drawing on their educational backgrounds, see opportunities to integrate interactivity with creative arts, suggesting that children could participate in activities like drawing or adding characters as part of the learning process. For instance, P15 advocated for incorporating art into English learning, stating, "It's not just about turning it into an English task but integrating it with different disciplines. Interdisciplinary education is trendy." Expanding on this, P15 proposed a multisensory approach to digital narratives, where children can influence the storyline by adding characters or actions in response to sensory cues. She explained, "You could add a layer where the child adds a character, and perhaps introduce the five senses. For example, if the oak leaf is happy and sings a song, what do you think the song sounds like? Could you sing a few lines?". P15 also described a scenario where characters faced emotional challenges and resolutions, prompting children to predict outcomes and create new characters with their parents. This not only enhances interactivity but also fosters engagement, creativity, and emotional ties between children and their parents. P13 emphasized the importance of play in learning, particularly for younger children. She advocated for integrating animation and playful elements into the learning process to sustain engagement. She explains, "It might involve turning it into something visual, like animation, constantly interacting with the kids or playing games with them, and learning while playing. For younger children, you really need to incorporate play and let them learn through it. If you just make them read books or picture books, they might only stay focused for 10 to 20 min before losing interest. So, engaging young children requires more effort, focusing on making activities fun and interactive."

5. Discussion

This study investigates the potential of an LLM-powered dialogue system to support young English Language Learners (ELLs) during e-book-mediated reading. Using Storio as a design probe, our research drew insights from children, parents, and educators. The examination of the child-agent interaction during a user study suggests that LLM-based CA simulates high-quality dialogues similar to those between

children and adults during reading. These capabilities suggest promise for enhancing language and cognitive outcomes for young learners, while revealing areas for improvement. Throughout this paper, we aim to contribute to the growing literature on child-AI interaction by offering empirical evidence and design insights on LLMs as children's conversational partners in educational settings. We discuss pedagogical and designimplications to inform future iterations of children's educational technology, highlight the study's limitations and outline directions for future research.

5.1. LLM-based CA as EFL children's dialogic reading partner

Prior research has examined the use of CAs as children's dialogic reading partners, but these systems have been predominantly rule-based, limiting their flexibility and capability to handle dynamic conversations, and leading to frequent misclassification and communication breakdowns. Our study stands out first for incorporating an embodied CA specifically designed for children's foreign language acquisition.

Through detailed observation and evidence-based analysis, we found that the LLM-based CA demonstrated significantly greater adaptivity than prior rule-based voice interfaces in literacy apps for children (Xu et al., 2021a). It provided differentiated feedback based on children's English proficiency offering elaboration to more proficient learners and simplified guidance to those needing more support. Its tolerance for speech-to-text errors helped reduce language anxiety by maintaining a positive and encouraging tone. For example, when a child's pronunciation was unclear, the CA responded with a supportive recast rather than an explicit correction, reinforcing the target structure while preserving conversational flow. This approach fosters supportive conversational dynamics and young learners' confidence and engagement in the reading activity. By contrast, a rule-based system would likely flag this as unrecognized or incorrect, interrupting the natural conversational flow. To further improve transparency without compromising encouragement, future designs might embed meta-prompts about automatic speech recognition (ASR) limitations (e.g., "I might mishear 'Little Oak'-let's try together"). Such cues can help set expectations, maintain trust, and support children's metacognitive awareness of the system's constraints.

In addition to expected scaffolding strategies, LLM-based CA also generated unanticipated yet pedagogically meaningful responses that supported child language acquisition. Its varied feedback enriched the child-agent interaction framework proposed by Xu et al. (2021a). Notably, the CA frequently employed corrective feedback techniques beneficial for second language learners (Roothooft and Breeze, 2016). These included translanguaging practices—repeating key lexical items from children's responses to reinforce meaning and promote retention (Daniel and Pacheco, 2016)—as well as recasts, a widely recognized implicit feedback strategy in L2 acquisition (Ellis et al., 2006). When children demonstrated consistent accuracy, the CA prompted additional target language use, fostering extended production. Such flexible and varied scaffolding is beyond the scope of previous preprogrammed rule-based systems. According to Xu (2023), CA is most effective when it aligns with children's cultural and linguistic practices Xu (2023). In the EFL context, code switching and translation are common practices during parent–child dialogic reading (Yang et al., 2022) and served as a discourse strategy in the classroom for effective instruction (Bhatti et al., 2018). Leveraging the LLM's multilingual capacity, future work should implement explicit translanguaging design policies, such as parallel phrasing, gentle code-switch prompts, learner-controlled toggles for target language proportion, to better reflect authentic EFL pedagogical discourse and accommodate diverse learner needs.

Consistent with previous studies, we observed that baseline English proficiency rather than age, could be a potential determinant of conversational success for EFL learners (Cheng et al., 2024). Future research could further explore this proficiency-based differentiation by diversifying prompt types and interactional structures, moving beyond comprehension questions alone (Xiao et al., 2025). Existing research (Cheng et al., 2024; Liu et al., 2022; Xu et al., 2021b) only embedded universal comprehension questions within the story for all learners. This leaves open the potential to investigate how different types of prompts might further support EFL children's engagement and language development across varying proficiency levels (Krathwohl, 2002). Repetitive practices embedded in chatbot design can offer EFL children useful affordances, especially lower-proficiency learners, but should be adopted strategically to avoid diminishing their engagement(Jeon et al., 2023; Xu et al., 2021b). Future work might also consider drawing on Bloom's taxonomy (Krathwohl, 2002) to develop differentiated prompt strategies that align with learners' cognitive and linguistic readiness, enabling more personalized and pedagogically grounded scaffolding.

Though improved compared to earlier voice recognition systems, persistent challenges with speech recognition for children, especially those with accented speech, highlight the need for further technical refinement. Enhancing the accuracy and responsiveness of LLM-based CAs is essential to support natural conversational flow and minimize frustration caused by miscommunication.

5.2. Leveraging contextual cues to support sense-making in language learning

Prior research on CAs as children's reading companions has largely focused on paper-based formats, where CAs prompt children with content-based questions tied to the story. For instance, Xu et al. (2021b) trained a CA to scaffold responses by shifting from open-ended to closed-ended questions when children encountered difficulties. While this strategy helps sustain interaction, it may inadvertently frame the reading experience as a task-oriented Q&A session, potentially undermining narrative engagement and deeper comprehension.

In contrast, digital learning environments offer multimodal affordances, such as rich visuals, animations and touch-based navigation that can be harnessed to aid meaning-making in foreign language learning. In EFL contexts, where learners often face heightened speaking anxiety and reduced linguistic confidence, a CA that guides multimodal sense-making can provide additional contextual cues, reducing their reliance on language alone to engage in the story and lower their affective filter (Morgan, 2013). A conversational agent that encourages children to draw on visual, narrative, and textual elements facilitates meaning-making through multiple semiotic resources, thereby reducing cognitive and linguistic load (Xiao et al., 2025).

Furthermore, this approach addresses concerns typically associated with multimedia learning environments. As Moreno and Mayer's (Moreno and Mayer, 2007) Cognitive-Affective Theory of Learning with Media (CATLM) suggests, when multimedia features are paired with appropriate instructional guidance through a pedagogical agent, they can enhance behavioral engagement and support active knowledge construction, leading to deeper cognitive engagement and learning. In our study, the conversational agent guided children to attend to relevant multimodal cues, thereby redirecting their focus to key content and mitigating the potential negative effects of multimedia on children's comprehension in multimedia-rich environments (Baker et al., 2001; Chow et al., 2010). Prior studies have similarly found that children feel more engaged and demonstrate better learning outcomes when interactions with the content are meaningful and contextualized(Liu et al., 2022). Building on this evidence, context-sensitive prompts delivered by a pedagogical agent can be strategically designed to leverage multimodal resources within the story. This fosters an interactive learning ecology in which learners actively construct meaning through coordinated engagement with the digital environment, which is an especially valuable strategy in EFL contexts where learners benefit from reduced linguistic demand and increased contextual support. Future research could investigate how prompt engineering can be used to generate context-sensitive dialogue in a stable and instructionally purposeful manner.

5.3. Diversifying emotion-responsive feedback channels

Feedback from parents, children, and educators suggested that the current voice of the LLM-powered conversational agent lacked emotional diversity and expressiveness. They noted that children learning English require significant emotional support, particularly when they make mistakes. Prior designs have largely embedded socio-affective cues in lexical choices (Liu et al., 2022), yet paralinguistic properties of speech strongly shape learners' perceptions of a CA's intelligence.Future designs should consider integrating more expressive language into CA, and introducing a range of emotional acoustic features (e.g., tones, prosody, pitch) and expressions in the CA's voice to better align with the emotional needs of children (Heerink et al., 2010). Nevertheless, in EFL setting, where speaking anxiety is common, voice alone is unlikely to suffice. Feedback mechanisms also influence engagement, and existing approaches that emphasize linguistic accuracy and limited visual motions do not fully address children's emotional requirements. To broaden the affective channel, the CA might incorporate additional social cues across modalities (visual and auditory). For example, animated facial expressions and brief, non-verbal sound cues that signal encouragement or empathy could strengthen emotional connection and support. A human-centred, multimodal CA that monitors children's facial expressions and vocal tone in real time could adapt feedback accordingly. In parallel, improving language models' handling of contextual nuance and memory recall can enhance perceived intelligence and, by extension, children's information uptake (Breazeal et al., 2016).

5.4. Support multimodal interaction within and beyond the screen

In our study, many children and parents expressed a desire for additional ways of interaction with the CA beyond the primary question-answering format. This feedback reflects a broader need for multimodal channels through which children can express themselves and engage with the story. Notably, several participants specifically asked to draw within the story, suggesting an enthusiasm for integrating creative, hands-on activities into the digital reading experience. They also indicated a clear interest in fostering physical and social connections with the CA beyond the screen. Future research might explore embodied instantiations (e.g., embedding the CA in plush toys) and seamlessly coupling of screen-based CA with paper book reading to scaffold playful and developmentally appropriate experiences for children (Garg et al., 2022).

In family settings, where children are typically surrounded by siblings and parents, we observed that siblings frequently wanted to participate in the reading sessions. To support this, prior work implemented family-oriented prompts designed to encourage child–parent interactions during dialogic reading (Xu et al., 2023). However, multi-user contexts presented technical challenges, including difficulties with background noise, overlapping speech, and the CA's difficulty tracking speakers and turns accurately.To address these limitations and broaden the scope of interaction, we advocate diversifying interaction channels that go beyond verbal exchanges. This could include tactile interfaces or gesture-based interactions that allow children to physically manipulate story elements. Such multimodal affordances can enrich learning, reduce barriers for shy or beginner learners, and make sessions more inclusive and engaging for all EFL learners.

5.5. Reimagining the CA's role to foster agency and social presence

In this study, the CA was designed to function as a parent figure, mirroring the experience children have with their caregivers. Prior research on CAs in dialogic reading had merely defined the role of these agents (Xu and Warschauer, 2020b; Xu et al., 2021b; Zhang et al., 2022), and the conversational style of CAs typically lacks distinctive personality traits. However, feedback from children, parents, and educators revealed a strong desire for a more child-friendly and engaging CA. In most existing designs, the conversation between the child and agent usually follows a question-answering pattern, in which CA plays an authoritative role that constrains children's agency to initiate questions. Future designs might consider allowing children to ask task-relevant questions back to CA while maintaining the conversation flow and topical relevance. This requires more sophisticated prompt engineering, fine-tuning or reinforcement learning (RL) (Bai et al., 2022). For example, Rafailov et al. (2024) trained an LLM by with human feedback to learn human preferences.

Moreover, future research might also explore a broader repertoire of CA roles to enhance dialogic reading. Existing research has confirmed that peer-to-peer interaction in the EFL setting can bolster confidence, especially among lower-proficiency learners (Pinter, 2007). Many children in our study expressed an interest in reading with peers, which could potentially help reduce language learning anxiety relative to an authoritative tutor. Incorporating roles that simulate peer interactions, alongside options to customize the CA's appearance, might increase children's motivation and engagement.

Additionally, carefully designed pedagogical agents added a layer of social presence to the digital learning platform, which is believed to facilitate active cognitive processing when learners perceive the digital learning experience as social interactions (Schneider et al., 2022). This has been reflected in children's perception and their interaction with the CA in our study. As suggested by our findings, allowing children to customize the CA's appearance or even design their own agent (e.g., dressing it up), which allows children to co-construct their learning experience, may further personalize learning and increase motivation. Future studies might investigate whether such personalization boosts learner agency and improves language outcomes.

Finally, as the agent's anthropomorphism is enhanced, at the same time, future studies should consider including transparent communication about the CA's capabilities, limitations, and role boundaries which is essential for setting appropriate expectations and supporting safe and effective learning.

5.6. Implications for integrating CA-supported reading into efl learning ecologies

In EFL contexts, daily exposure to English is limited at both school and home. In such settings, additional, high-quality input beyond classroom time can significantly shape language learning outcomes (Chow et al., 2008, 2010). Systems like these can extend that exposure by supporting dialogic reading and practice at home; however, successful adoption requires coordinated, multi-stakeholder effort across teachers, caregivers, and developers rather than a purely "plug-and-play" approach. This study offers design implications from multiple stakeholder perspectives, but does not fully integrate these into design iterations. Future work should intentionally integrate these voices into the design process, creating systems that balance learner needs with classroom realities, build home–school bridges, and incorporate lightweight evaluation loops and analytics for rapid refinement (Li et al., 2025). Given the flexibility and accessibility of LLMs, educators and parents should be invited as co-designers of local implementations, drawing on first-hand knowledge of their children to configure interaction modalities and feedback routines.

In the present study, children used the system independently. While effective, the social presence of a parent may help sustain affective engagement, depending on parenting style and family dynamics (Dexter and Stacks, 2014). The system also provides a broadly accessible option when shared parent–child reading is infeasible or not preferred. In addition, CA-supported reading can be deployed in parent–child dyads to scaffold participation for caregivers with limited English proficiency, and in classrooms to supplement instruction when teacher support is constrained. Future studies should explore pathways for better integration of this kind of system across home and school settings.

Ethical consideration A primary concern with language learning models (LLMs) is the potential for generating biased or inappropriate

content. This risk is particularly salient when engaging with children, who may pose unpredictable or sensitive questions. Given LLMs are trained on large-scale internet corpora, there's a risk that children could be exposed to unsuitable content, negatively impacting their language learning experience and emotional well-being. To mitigate this, our application utilized prompt engineering to constrain the conversation's scope. Our prompts are specifically designed to evaluate the correctness of the student's answers and provide adaptive feedback. This approach helps prevent students from asking unrelated questions, but it also limits the application's versatility and restricts its interactive potential. For instance, if a student is curious about a new word from the story or wants to learn about the author's background, the CA might not be able to accommodate these queries within the current system. Future systems could include integrating a safety monitoring module, such as a separate safety control agent, to moderate outputs in real time. This would preserve the educational potential of open-ended interactions while ensuring age-appropriate content. Such improvements would also support more naturalistic, learner-driven dialogue, allowing CAs to better serve as adaptive and safe learning companions in diverse EFL contexts.

Accountability and responsibility Unlike human educators, LLM-based systems currently lack mechanisms for accountability, making it difficult to trace responsibility when errors occur in content delivery or interaction. In our study, mistakes made by Mia, our LLM-based agent, underscore the importance of supervision and continuous model development in future research. One potential solution is to implement active learning to improve reliability during future LLM fine-tuning stages. By systematically identifying instances where the model produces erroneous or suboptimal feedback, human experts can label corrective responses. These curated data can then be used for targeted fine-tuning, improving the reliability and pedagogical quality of the system over time. Such an approach supports iterative refinement and introduces a layer of human-in-the-loop accountability, essential for deploying AI systems for minorities.

6. Limitations

First of all, while this study adopted a well-established coding framework to analyze the quality of child-agent conversations, it does not directly compare them to real parent–child dialogic reading practices. As a result, it remains unclear how the CA's scaffolding compares to the high-quality interactions between parents and children. However, it is important to note that not all parent–child dialogues are of high quality. Future studies should take this variability into account when defining what constitutes high-quality child-agent conversations in comparison to human interactions.

Secondly, although most children answered post-reading comprehension questions correctly, their baseline English proficiency was reported by parents, which may not provide an objective assessment. Future research should consider using established language proficiency measures to evaluate children's baseline English skills and then conduct an intervention to formally assess both cognitive and affective learning gains.

Third, this study involved only short-term interaction with a single narrative story. Our study indicated that children were highly engaged, which could be attributed to the novelty of the system, as nearly all participants had not used such an application before. Furthermore, it remains unclear whether similar engagement and learning effects would persist across different story genres or during extended use. Longitudinal studies conducted in naturalistic, home-based environments would offer valuable insights into the sustainability of engagement, generalization of learning across content types, and long-term learning outcomes.

Moreover, although this study is situated in an EFL context, it focused on learners from a relatively homogeneous linguistic and cultural background, as all participants were recruited through snowball sampling. Given the growing interest in multilingual and multicultural learning environments, future research should build on this design and explore how CA-supported reading performs in diverse language communities, particularly among learners from varied socioeconomic backgrounds. Additionally, long-term interventions could examine how prompt engineering and system personalization can stably and intentionally support learners with varied linguistic profiles.

Finally, while this study employed in-depth coding of chat log data triangulated by multiple data sources, the current system did not incorporate multimodal learning analytics such as clickstream data, interaction timing, or behavioral patterns within the system. Future work could benefit from integrating richer data streams—such as navigation behaviors, response latencies, or eye-tracking measures—to generate more comprehensive understanding of children's engagement and learning.

7. Conclusion

This study presents Storio, an e-book integrated with an LLM-based bilingual conversational agent, designed to support EFL children's English reading and enhance their language learning outcomes. Using Storio as a design probe, we investigated how children interact with the agent during digital story reading and collected perspectives from parents and educators on its pedagogical value. Our findings suggest that the LLM-based agent can effectively simulate high-quality dialogic reading practices, providing scaffolding that enhances children's comprehension and engagement during reading sessions. This study offers valuable design implications from the perspectives of multiple stakeholders and demonstrates how LLM-based CAs can be leveraged to support children's language learning.

Statements on open data and ethics

The study was approved by the Institutional Review Board (IRB) at The Pennsylvania State University (Study ID: STUDY00024301). Informed consent was obtained from all participating children and their legal guardians. All privacy rights were strictly observed, and participation was voluntary. The data supporting the findings of this study are not publicly available due to privacy and ethical restrictions. However, data can be obtained upon reasonable request by contacting the corresponding author.

CRediT authorship contribution statement

Feiwen Xiao: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zhaohui Li:** Writing – original draft, Software, Investigation, Conceptualization. **Jiaju Lin:** Writing – original draft, Visualization, Software, Formal analysis, Data curation. **Xiaohan Zou:** Writing – original draft, Visualization, Software. **Dr Dandan Yang:** Software, Conceptualization. **Dr Ellen Wenting Zou:** Writing – review & editing, Supervision. **Dr Jinjun Xiong:** Writing – review & editing, Supervision.

Funding statement

This work is supported, in part, by the Award # 2229873 (National AI Institute for Exceptional Education) sponsored by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education, and Grant # R305C240046 (Center for Early Literacy and Responsible AI) sponsored by Institute of Education Sciences, U.S. Department of Education. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Tables

See Table A6 and Table A7.

Table A6
Coding schema for children's responses and CA's feedback.

Coding Item	Definition	Example
Correct	children answered the CA's question accurately and relevantly	CA: Why are the butterflies so popular? CHI: Because they have many colors.
Incorrect	incorrect, irrelevant answer, or indicate "I don't know"	CA: What does Little Oak look like in the water? CHI: I tried it!
Partial Correct	contained elements of correctness, but were either unfinished or grammatically incorrect	CA: What colors are the wings of the butterflies? And what do they look like? CHI: Rainbow.
Unclear	response not correctly identified by the CA	CHI: The black bird 班 on the little OK is bad Yuk CHI: Because the liquor oak is like a dry leaf.
Contingency	feedback specifically addressed the child's response logically instead of merely providing generic or unrelated feedback	CA: There are four guards on the way to save the butterfly friends. Who are the four guards? CHI: A frog or lizard. CA: Good try, but there's more than two guards. Can you remember who else was there? CHI: A snake and a black bird.
Extension	feedback contained additional information instead of merely verifying the child's response.	CA: Why did not the guards see the Little Oak? CHI: Because they saw the little old is a dry leaf. CA: Well done! Little Oak used his oak color to camouflage himself as a dry leaf!
Encouragement	motivated the child to engage further in the conversation	CA: Here comes the first question. What does Little Oak look like in the water? CHI: I don't know. CA: You can do it! Look again, how did Little Oak see himself in the water?
Explicit Correction	CA explicitly points out mistakes in grammar, vocabulary, and meaning	CA: Here comes the first question. What does Little Oak look like in the water? CHI: He's a tree. CA: Good try, but remember Little Oak sees himself as something else, not a tree. CHI: 一个枯叶(a dry leaf) CA: Good job! You're right. Little Oak looks like a dry leaf ("一个枯叶") in the water! Keep going!
Question Simplification	simplifies complex questions or concepts by providing more accessible explanations	CA: What do the other butterflies feel at this moment? CHI: They are very sad. CA: Hmm, think again. Little Oak just saved them. How do you think they feel about it? CHI: They are very happy.
Contextual Cues	references the illustrations or story context to help guide the child	CA: There are four guards on the way to save the butterfly friends. Who are the four guards? CHI: the blackbirds. CA: Oops! Let's try again. Maybe take a look again at the story where he went to save his friends. There are animals guarding!
Bilingual Engagement	promoting language switch	CA: Why are the butterflies so popular? CHI: 因为跳舞的butterfly (because dancing butterflies are beautiful) 很美丽 CA: Good job, but remember we're trying to answer in English. Try again, perhaps start with "Because the…" CA: What do the other butterflies feel at this moment? CHI: 高兴 (Happy) CA: Well done, that's right! The word '高兴' in English means 'happy'. The butterflies are indeed happy because they were saved by Little Oak.

Table A7
Outstanding examples of child-agent interaction in terms of language production, cognitive engagement, and affect.

Category	Example
Language production	When answering questions, "Who are the four guards?". Child P11 replied, "The guard is a frog (守卫是一只青蛙), a lizard (一只蜥蜴), a snake (一只蛇), a big black bird (一只大黑鸟)." Before this, he had attempted to answer twice but failed to be recognized by the agent because the words "lizard" and "guard" were new to him, which prolonged his response due to undeveloped fluency.
Cognitive Engagement	**High Proficiency:** Child (P10) exhibited the ability to apply newly learned vocabulary in context. When asked, "What is the blackbird doing when Little Oak arrived at his backyard?", she initially used the phrase "taking a nap", and then adaptively reformulated it as "napping" in a follow-up sentence. This flexible use of language suggests emerging control over morphological forms and reflects both cognitive and linguistic agility, which is an indicator of meaningful vocabulary uptake and transfer in context. **Low proficiency:** Child (P11) struggled with the pronunciation of "popular" and received no recognition from the CA, he questioned his accuracy by asking, "Am I right?", before using multimedia tools to aid his understanding. This child hild independently utilized multimedia affordances embedded in the application. He repeatedly clicked on the vocabulary and narration features to hear the correct pronunciation. He then engaged in self-directed practice by reading aloud multiple times. This episode illustrates how lower-profienct children effectively using the CA as a tool for pronunciation practice and vocabulary reinforcement.
Affect	**Older child:** An amusing yet telling incident involved a misrecognition error where the agent confused "look like a dry leaf (像枯叶)" with "leaf of Shiitake Mushroom (香菇叶)," prompting discreet laughter from the child, indicative of an engagement with and cognitive recognition of the CA's limitations. **Younger child:** Child (P12) enthusiastically responded, "I am ready," followed by a smile and visible excitement as she leaned back on the couch. Her joy was further amplified when she correctly answered a question after four turns of conversation. A smiling face appeared beside the CA on the screen, prompting her to point at the smiling face and laugh. Similarly, another younger participant (P7) responded to the question, "Why does the blackbird catch all of the butterflies except for Little Oak?" with laughter, playfully singing the answer in Chinese, "Because Little Oak has camouflage."

Data availability

The data used in this study is not open to the public.

References

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.

Baker, L., Mackler, K., Sonnenschein, S., & Serpell, R. (2001). Parents' interactions with their first-grade children during storybook reading and relations with subsequent home reading activity and reading achievement. Journal of School Psychology, 39(5), 415–438.

Bhattacharyya, S. (2021). A brief idea on the importance of first-generation English learners in the contemporary universe. International Journal of English Learning & Teaching Skills, 3(2), 2097–2100.

Bhatti, A., Shamsudin, S., & Said, S. B. M. (2018). Code-switching: a useful foreign language teaching tool in EFL classrooms. English Language Teaching, 11(6), 93–101.

Boehner, K., Vertesi, J., Sengers, P., & Dourish, P. (2007). How HCI interprets the probes. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 1077–1086).

Breazeal, C., Harris, P. L., DeSteno, D., Kory Westlund, J. M., Dickens, L., & Jeong, S. (2016). Young children treat robots as informants. Topics in Cognitive Science, 8(2), 481–491.

Brown, T. B. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

Caesar, L. G., & Nelson, N. W. (2014). Parental involvement in language and literacy acquisition: a bilingual journaling approach. Child Language Teaching and Therapy, 30(3), 317–336.

Cheng, X., Yin, L., Lin, C., Shi, Z., Zheng, H.Zhu, L. …Dong, R. (2024). Chatbot dialogic reading boosts comprehension for Chinese kindergarteners with higher language skills. Journal of Experimental Child Psychology, 240, 105842.

Cheng, Y., Yen, K., Chen, Y., Chen, S., & Hiniker, A. (2018). Why doesn't it work? Voice-driven interfaces and young children's communication repair strategies. In Proceedings of the 17th ACM conference on interaction design and children (pp. 337–348).

Chow, B. W.-Y., McBride-Chang, C., Cheung, H., & Chow, C. S.-L. (2008). Dialogic reading and morphology training in Chinese children: Effects on language and literacy. Developmental Psychology, 44(1), 233.

Chow, B. W.-Y., McBride-Chang, C., & Cheung, H. (2010). Parent–child reading in English as a second language: effects on language and literacy development of Chinese kindergarteners. Journal of Research in Reading, 33(3), 284–301.

Daniel, S. M., & Pacheco, M. B. (2016). Translanguaging practices and perspectives of four multilingual teens. Journal of Adolescent & Adult Literacy, 59(6), 653–663.

DesJardin, J. L., Stika, C. J., Eisenberg, L. S., Johnson, K. C., Ganguly, D. M. H., Henning, S. C., & Colson, B. G. (2017). A longitudinal investigation of the home literacy environment and shared book reading in young children with hearing loss. Ear and Hearing, 38(4), 441–454.

Dexter, C. A., & Stacks, A. M. (2014). A preliminary investigation of the relationship between parenting, parent-child shared reading practices, and child development in low-income families. Journal of Research in Childhood Education, 28(3), 394–410.

Dietz, G., Le, J. K., Tamer, N., Han, J., Gweon, H., Murnane, E. L., & Landay, J. A. (2021). Storycoder: Teaching computational thinking concepts through storytelling in a voice-guided app for children. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1–15).

Dijkstra, R., Genç, Z., Kayal, S., Kamps, J., et al. (2022). Reading comprehension quiz generation using generative pre-trained transformers. In ITextbooks@ AIED (pp. 4–17).

Dong, Y., & Chow, B. W.-Y. (2022). Home literacy environment and English as a second language acquisition: a meta-analysis. Language Learning and Development, 18(4), 485–499.

Druga, S., Williams, R., Park, H. W., & Breazeal, C. (2018). How smart are the smart toys? Children and parents' agent interaction and intelligence attribution. In Proceedings of the 17th ACM conference on interaction design and children, IDC '18 (pp. 231–240). New York, NY, USA: Association for Computing Machinery. ISBN https://doi.org/10.1145/3202185.3202741 9781450351522.

Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. Studies in second language acquisition, 28(2), 339–368.

Ergül, C., Akoğlu, G., Karaman, G., & Sarıca, A. D. (2017). Effects of dialogic reading program on later reading skills: Follow-up study. Journal of Theoretical Educational Science, 10(2), 191–219.

Fan, Y., Jiang, F., Li, P., & Li, H. (2023). Grammargpt: exploring open-source LLMs for native Chinese grammatical error correction with supervised fine-tuning. In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 69–80). Springer.

Fitton, L., McIlraith, A. L., & Wood, C. L. (2018). Shared book reading interventions with English learners: a meta-analysis. Review of Educational Research, 88(5), 712–751.

Flynn, K. S. (2011). Developing children's oral language skills through dialogic reading: guidelines for implementation. Teaching Exceptional Children, 44(2), 8–16.

Garg, R., Cui, H., Seligson, S., Zhang, B., Porcheron, M.Clark, L. …Beneteau, E. (2022). The last decade of HCI research on children and voice-based conversational agents. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1–19).

Goldenberg, C. (2008). Teaching English language learners: what the research does-and does not-say.

Hargrave, A. C., & Sénéchal, M. (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), 75–90.

He, D., & Li, D. C. S. (2021). Glocalizing ELT reform in China: a perspective from the use of English in the workplace. RELC Journal, 54, 149–165. https://doi.org/10.1177/00336882211018499

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Relating conversational expressiveness to social presence and acceptance of an assistive social robot. Virtual Reality, 14, 77–84.

Heritage, J. (2004). 5 conversation analysis and institutional talk. In Handbook of language and social interaction (pp. 103–147). Psychology Press.

Hsieh, C.-P., Sun, S., Kriman, S., Acharya, S., Rekesh, D., Jia, F., & Ginsburg, B. (2024). Ruler: What's the real context size of your long-context language models? arXiv preprint arXiv:2404.06654.

Hsieh, Y., & Huang, S. (2020). Using an e-Book in the secondary English classroom: effects on EFL reading and listening. Education and Information Technologies, 25(2), 1285–1301.

Jeon, J., Lee, S., & Choe, H. (2023). Beyond ChatGPT: a conceptual framework and systematic review of speech-recognition chatbots for language learning. Computers & Education, 206, 104898.

Jiménez-Castellanos, O., & García, E. (2017). Intersection of language, class, ethnicity, and policy: toward disrupting inequality for English language learners. Review of Research in Education, 41(1), 428–452.
Jin, J. (2005). Which is better in China, a local or a native English-speaking teacher? English Today, 21(3), 39–46.
Kasper, G., & Wagner, J. (2014). Conversation analysis in applied linguistics. Annual Review of Applied Linguistics, 34, 171–212.
Kennedy, C., & McLoughlin, A. (2023). Developing the emergent literacy skills of English language learners through dialogic reading: a systematic review. Early Childhood Education Journal, 51(2), 317–332.
Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. Advances in Neural Information Processing Systems, 35, 22199–22213.
Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: an overview. Theory into Practice, 41(4), 212–218.
Lever, R., & Sénéchal, M. (2011). Discussing stories: on how a dialogic reading intervention improves kindergartners' oral narrative construction. Journal of Experimental Child Psychology, 108(1), 1–24.
Li, Y., Qu, S., Shen, J., Min, S., & Yu, Z. (2023). Curriculum-driven edubot: a framework for developing language learning chatbots through synthesizing conversational data. arXiv preprint arXiv:2309.16804.
Li, Z., Xiao, F., Lin, J., Zou, X., Zheng, Q., & Xiong, J. (2025). Storylab: empowering personalized learning for children through teacher-guided multimodal story generation. In International conference on artificial intelligence in education (pp. 285–292). Springer.
Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z.Gu, H. …Tao, J. (2024). Gpt-4v with emotion: a zero-shot benchmark for generalized emotion recognition. Information Fusion, 108, 102367.
Liu, C.-C., Liao, M.-G., Chang, C.-H., & Lin, H.-M. (2022). An analysis of children' interaction with an AI chatbot and its impact on their interest in reading. Computers & Education, 189, 104576.
Lonigan, C. J., & Whitehurst, G. J. (1998). Relative efficacy of parent and teacher involvement in a shared-reading intervention for preschool children from low-income backgrounds. Early Childhood Research Quarterly, 13(2), 263–290.
Lovato, S., & Piper, A. M. (2015). "Siri, is this you?" understanding young children's interactions with voice input systems. In Proceedings of the 14th international conference on interaction design and children (pp. 335–338).
MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. Language Learning, 44(2), 283–305.
Meyer, J., Jansen, T., Schiller, R., Liebenow, L. W., Steinbach, M., Horbach, A., & Fleckenstein, J. (2024). Using llms to bring evidence-based feedback into the classroom: AI-generated feedback increases secondary students' text revision, motivation, and positive emotions. Computers and Education: artificial Intelligence, 6, 100199.
Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. Scientific Reports, 14(1), 9603.
Mol, S. E., Bus, A. G., De Jong, M. T., & Smeets, D. J. H. (2008). Added value of dialogic parent–child book readings: a meta-analysis. Early Education and Development, 19(1), 7–26.
Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments: special issue on interactive learning environments: contemporary issues and trends. Educational Psychology Review, 19(3), 309–326.
Morgan, H. (2013). Multimodal children's e-books help young learners in reading. Early Childhood Education Journal, 41, 477–483.
Nakamura, S., Phung, L., & Reinders, H. (2021). The effect of learner choice on L2 task engagement. Studies in Second Language Acquisition, 43(2), 428–441.
Perez, E., Kiela, D., & Cho, K. (2021). True few-shot learning with language models. Advances in Neural Information Processing Systems, 34, 11054–11070.
Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. Annual Review of Applied Linguistics, 36, 50–72.
Pinter, A. (2007). Some benefits of peer–peer interaction: 10-year-old children practising with a communication task. Language Teaching Research, 11(2), 189–207.
Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2024). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36.
Roothooft, H., & Breeze, R. (2016). A comparison of EFL teachers' and students' attitudes to oral corrective feedback. Language Awareness, 25(4), 318–335.
Ruusuvuori, J. (2012). Emotion, affect and conversation. The Handbook of Conversation Analysis, 330–349.
Sabharwal, N., & Agrawal, A. (2020). Cognitive virtual assistants using Google Dialogflow: Develop complex cognitive bots using the Google Dialogflow platform. Apress.
Schneider, S., Beege, M., Nebel, S., Schnaubert, L., & Rey, G. D. (2022). The cognitive-affective-social theory of learning in digital environments (CASTLE). Educational Psychology Review, 34(1), 1–38.
Seo, Y. (2023). The role of home language environment and parental efforts in children's English development in an EFL context. Journal of Multilingual and Multicultural Development, 1–15.
Van Brummelen, J., Tian, M. C., Kelleher, M., & Nguyen, N. H. (2023). Learning affects trust: design recommendations and concepts for teaching children—and nearly anyone—about conversational agents. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, pp. 15860–15868).
Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al (2022a). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.
Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al (2022b). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.
Whitehurst, G. J., Falco, F. L., Lonigan, C. J., Fischel, J. E., DeBaryshe, B. D., Valdez-Menchaca, M. C., & Caulfield, M. (1988). Accelerating language development through picture book reading. Developmental Psychology, 24(4), 552.
Xiao, F., Zhao, P., Sha, H., Yang, D., & Warschauer, M. (2023). Conversational agents in language learning. Journal of China Computer-Assisted Language Learning.
Xiao, F., Zou, E.W., Lin, J., Li, Z., & Yang, D. (2025). Parent-led vs. AI-guided dialogic reading: Evidence from a randomized controlled trial in children's e-book context. British Journal of Educational Technology, 56(5), 1784–1813.
Xu, Y. (2023). Talking with machines: can conversational technologies serve as children's social partners? Child Development Perspectives, 17(1), 53–58.
Xu, Y., & Warschauer, M. (2020a). Exploring young children's engagement in joint reading with a conversational agent. In Proceedings of the Interaction Design and Children Conference (pp. 216–228).
Xu, Y., & Warschauer, M. (2020b). 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 (pp. 1–13).
Xu, Y., Branham, S., Deng, X., Collins, P., & Warschauer, M. (2021a). Are current voice interfaces designed to support children's language development? In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1–12).
Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021b). Same benefits, different communication patterns: comparing children's reading with a conversational agent vs. A human partner. Computers & Education, 161, 104059.
Xu, Y., He, K., Vigil, V., Ojeda-Ramirez, S., Liu, X.Levine, J. …Warschauer, M. (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 (pp. 160–172).
Yang, D., Xia, C., Collins, P., & Warschauer, M. (2022). The role of bilingual discussion prompts in shared e-Book reading. Computers & Education, 190, 104622.
Yeung, S. S., & King, R. B. (2016). Home literacy environment and English language and literacy skills among Chinese young children who learn English as a second language. Reading Psychology, 37(1), 92–120.
Zevenbergen, A. A., & Whitehurst, G. J. (2003). Dialogic reading: a shared picture book reading intervention for preschoolers. On Reading Books to Children: parents and Teachers, 177, 200.
Zhang, C., Liu, X., Ziska, K., Jeon, S., Yu, C.-L., & Xu, Y. (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 (pp. 1–23).
Zhang, Z., Xu, Y., Wang, Y., Yao, B., Ritchie, D.Wu, T., Li, T.-J.-J. (2022). Storybuddy: a human-AI collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In Proceedings of the 2022 CHI conference on human factors in computing systems (pp. 1–21).
Zhang, Z., & Huang, X. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. Heliyon, 10(3).
Zou, W., Yang, Q., DiFranzo, D., Chen, M., Hui, W., & Bazarova, N. N. (2024). Social media co-pilot: designing a chatbot with teens and educators to combat cyberbullying. International Journal of Child-Computer Interaction, 100680.