PolyTaler: A Multi-Modal Interactive AI Agent to Enhance Children's Reading Engagement

ANONYMOUS AUTHOR(S)

Reading engagement remains a critical challenge for children, with existing digital reading support technologies often failing to address children's multi-modal communication patterns and individual learning needs. We introduce PolyTaler, a conversational, multi-modal reading companion for elementary school children. Guided by formative interviews with 16 parents, PolyTaler addresses four gaps in current systems: insufficient multi-modal emotional support, lack of engaging visual cues to sustain attention, limited adaptivity for deeper thinking, and inadequate parental involvement. The system senses children's affect from both voice and text, creates supportive illustrations, guides a three-step dialogue from understanding to reasoning to personal connection, and provides a dashboard supporting parental interaction. Our evaluation with 12 child–parent pairs showed higher interest and sustained focus with PolyTaler, while parents reported that the system prompted deeper thinking and fostered parent–child interaction. This work advances reading agent design principles and shows AI's promise in enriching home reading.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; • **Social and professional topics** → **Children**.

Additional Key Words and Phrases: human-AI interaction, children-agent interaction, multi-modal, reading, education

ACM Reference Format:

Anonymous Author(s). 2018. PolyTaler: A Multi-Modal Interactive AI Agent to Enhance Children's Reading Engagement. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 23 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Reading plays a crucial role in children's cognitive, emotional, and social development, forming the foundation for academic achievement and lifelong learning [22, 24]. Beyond basic literacy skills, engagement with literature enhances children's emotional intelligence, builds critical thinking, and develops their capacity to understand others' perspectives [23]. These benefits are especially notable for elementary-age readers who engage with rich narratives featuring nuanced characters and meaningful situations [17, 42]. However, such texts can be demanding. Children may struggle with unfamiliar vocabulary, complex syntax, or a lack of background knowledge [46, 52]. These cognitive challenges can, in turn, lead to wavering attention and diminished motivation, especially when repeated difficulty erodes self-efficacy or when the material feels disconnected from personal interests [9, 26]. While joint reading with a more knowledgeable partner is a powerful strategy, its real-world implementation is often hindered by practical barriers [7, 66]. Research has shown that during the joint reading process, the partner may miss opportune moments for discussion [72], lack effective prompts that elicit deeper talk [60], or face time constraints that limit consistent, high-quality support [66]. These barriers point to the need for a more effective assistance that can scaffold comprehension, sustain attention, and strengthen motivation during children's reading.

Manuscript submitted to ACM

Researchers have developed conversational agents (CAs) as reading partners to support children's reading, with systems evolving from basic text-based interactions to voice-enabled agents that more closely resemble human-guided reading. Representative systems include *StoryMate*, which features LLM to support young children's story-reading [6]; **CharacterCritique**, which uses analytical stories as the medium to promote critical thinking of children [55]; *StoryBuddy*, which supports human–AI co-reading with flexible parental involvement [74]; *Elinor*, which integrates conversational AI into children's educational programming [63]; and *TaleMate*, which facilitates parent–child joint reading through voice interaction [52]. Empirical studies demonstrate that these agents can increase narrative-relevant talk, reduce off-task vocalizations, and improve story comprehension [59], in some cases achieving effects comparable to human partners [64]. In the home, families weave voice interfaces into bedtime and after-school routines, enabling children to request stories or definitions on demand [62, 74, 75]. At the same time, parents leverage the agent's prompts to keep sessions flowing and share reading responsibilities [60, 66].

Despite these promising advances, current systems are constrained by weak support for multi-modal interaction and thus have several limitations. First, many CAs rely on a single channel of interaction, which cannot capture children's natural signals such as expressions and vocal patterns to signal children's comprehension, uncertainty, and affect [27, 46, 52, 66]. Moreover, emotional awareness is limited: agents rarely detect whether a child is engaged, frustrated, or bored, and they seldom adjust their questions or pacing in response [6, 24, 54, 63]. Additionally, personalization remains shallow; many systems depend on pre-programmed dialogue flows or fixed reading materials, resulting in one-size-fits-all scripts that fail to accommodate differences in reading level, attention span, or learning preferences [9, 54, 74].

Our research aims to address these limitations by designing and implementing an enhanced multi-modal AI agent for joint reading with elementary school children (ages 6-10) to improve both reading engagement and efficiency. We define "engagement" as the child's active, sustained, and emotionally invested participation in reading [13, 33], whereas "reading efficiency" is the learning gained from reading, including improved comprehension and inferential reasoning rather than mere speed [30, 43]. To understand the challenges in children's reading behaviors and gather parental perspectives on design considerations, we conducted 60-minute semi-structured interviews and parent-centered contextual design sessions with 16 parents of children ages 6 to 10. We adopted a parent-centered approach because parents orchestrate home reading and can reliably report engagement and routines in context. Children aged 6–10 seldom articulate these insights, and prior studies have successfully relied on parents [60, 66, 72, 74, 75]. Our findings show that parents observed stronger engagement when children chose their own books and set simple goals. Visuals were most effective as targeted scaffolds for abstract or difficult content, though confident readers often preferred fewer images. Parents also emphasized the need for personalization that adapts pacing, question depth, and feedback to children's ability, attention, language strengths, and fatigue. With regard to children's privacy, parents preferred emotion detection based on voice and text that runs locally on the device. For parental involvement, parents described their role as coaching rather than co-reading, and favored concise discussion highlights and simple progress signals over right–wrong checks.

These results inform four design considerations (DCs): DC1 adopts on-device emotion detection from voice and text to guide pacing and encouragement rather than grading; DC2 provides supportive illustrations that clarify abstract content and avoid distracting through interaction design; DC3 structures a three-step dialogue that moves from understanding to reasoning to personal connection with adaptive difficulties; and DC4 offers a parent dashboard that summarizes themes, key questions, and salient moments to support reflective conversations at home.

Building on these insights, we designed and implemented PolyTaler, an AI agent system that leverages multi-modal inputs to support interactive storytelling with children. Guided by our key insights and design considerations, PolyTaler integrates three core components: (1) a visual cue generator that activates schema knowledge and contextualizes upcoming content; (2) an emotion detection module that captures multi-modal affective signals to guide adaptive responses; and (3) a personalized multi-agent framework, organized around a three-stage interaction model, that dynamically generates questions and responses in adaptation to children's engagement and emotions. In addition, PolyTaler tracks children's interactions and progress, presenting performance data in a parental dashboard that enables caregivers to assess comprehension development.

To understand how children perform with the assistance of PolyTaler and how parents perceive our system, we conducted a counterbalanced mixed-methods user evaluation with 12 child-parent pairs. Each child read chapters with and without PolyTaler; we collected comprehension scores, ratings of interest and attention, interaction logs, and interviews. Quantitatively, reading with PolyTaler increased children's interest and sustained attention, while comprehension gains were modest and variable. Qualitatively, three themes stood out: brief, well-timed questions helped children stay on task and make connections; simple illustrations shown before and after chapters supported understanding without interrupting reading; and parents preferred a dashboard that offered concise conversation starters for follow-up rather than test-style scores. Overall, the study indicates that a conversational, multi-modal companion can enrich home reading by elevating engagement while respecting parent-child roles.

This paper makes the following main contributions:

- We conducted a formative investigation with children and parents, revealing key design needs for multi-modal reading support: affective sensing through voice and text, visual cues to sustain engagement, a structured three-step dialogue to foster deeper thinking, and a parental dashboard to scaffold follow-up interactions.
- We developed PolyTaler, a multi-modal interactive AI system that integrates a visual cue generator, an emotion detection module, and a personalized multi-agent workflow, complemented by a parent-facing dashboard for reflective engagement.
- We evaluated the efficacy of PolyTaler through a controlled user study combining quantitative and qualitative methods, demonstrating its impact on children's engagement, comprehension, and child-agent collaboration during reading activities.

2 Related Work

2.1 Children's Reading Practices and Joint Reading

Literary reading supports children's cognitive, emotional, and social development by strengthening theory of mind, language growth, and critical interpretation [22–24, 42]. Engagement with linguistically rich texts, including age-appropriate adaptations of classics, expands vocabulary and exposes learners to complex syntax; when text difficulty exceeds independent reading level, listening to complex material can still yield measurable vocabulary gains [24, 50]. These benefits depend on high-quality adult–child interaction around texts, where attentive listening and contingent talk validate and extend children's contributions [22].

However, children's reading is often hindered by persistent obstacles such as unfamiliar vocabulary, complex syntax, and knowledge gaps that undermine both comprehension and motivation [46, 52]. Given that attention fluctuates and learners are heterogeneous, uniform instruction is often inadequate. This is supported by longitudinal studies showing

varied developmental trajectories, which highlights the need for personalized scaffolding over a fixed educational pathway [9, 26].

Joint reading with a more knowledgeable partner addresses these challenges with consistent empirical support. Research showed that during joint reading, dialogic techniques deliver reliable vocabulary gains of about 2.6 to 3.0 words per session, with an additional gain of about 1.2 words when dialogic prompts are used [15, 34]. Strategic questioning and scaffolded feedback deepen comprehension [19, 71], open questions elicit longer and more complex child language [11], explicit vocabulary support accelerates word learning [24], and discussion of characters' emotions strengthens emotional understanding [5]; appropriately used humor can help sustain attention [7, 15, 49]. However, its implementation remains a barrier in practice: these reading partners often struggle to locate useful pause points and sustain rich dialogue, time and skill constraints limit access, and effects attenuate for at-risk populations [34, 60, 66, 72], which underscores the need for accessible, consistent, adaptive supports.

2.2 Children–AI Interaction in Educational Contexts

Children interact with technology in multi-modal ways that integrate gesture, gaze, facial expression, and prosody, and younger children often attribute social qualities to AI systems [18, 46, 60, 67]. Older children are more likely to probe capabilities and limits, yet they still expect systems to accommodate rich communicative signals [18, 58]. These tendencies create opportunities for engagement but challenge systems that primarily process verbal input and overlook non-verbal cues central to children's expression and self-regulation [60].

The current educational AI landscape is broad, spanning intelligent tutors, adaptive platforms, creative co-design tools, and voice-enabled applications that structure activity, deliver feedback, and adjust difficulty [8, 31, 45, 51, 57, 68, 73]. However, a closer look at existing systems, particularly voice-based ones, reveals significant limitations. Analytic evaluations of children's voice apps show that most prompts are restrictive and that fewer than six in ten apps extend responses to sustain engagement, which limits back-and-forth dialogue and children's language production [60]. In-home studies document how preschoolers ask questions of commercial assistants and where answer quality and turn-taking break down; they also show how shared, always-available devices reshape family routines and parenting practices by democratizing access [1, 27]. For multilingual families, early design investigations motivate conversational objectives that support bilingual acquisition and inclusive participation [2]. HCI work further cautions that tightly scripted voice apps can create pseudo-interactions that feel interactive yet fail to support authentic learning [10].

These documented shortcomings point to three recurring, fundamental gaps in current child-AI interaction. First, a non-verbal gap persists, as systems rarely interpret children's gestures and other non-verbal cues, despite evidence that children naturally use these channels even with voice-only interfaces[60, 72]. Second, emotional responsiveness is limited, although affect shapes motivation and outcomes; studies of reading robots and social agents report excitement, disappointment, and efforts to share personal feelings, which suggests the value of emotionally contingent responses and social connectedness [4, 48, 70]. Third, adaptivity is often restricted to performance-based adjustments; yet children differ markedly in attention, regulation, and learning approaches, and cluster analyses reveal distinct interaction profiles with divergent outcomes under identical tools [9, 26, 54].

Addressing these gaps requires moving towards hybrid human–AI frameworks that practice conditional automation, aligning with HCI principles of co-regulation, transparency, and child-centred design[14, 20, 35, 36, 53].

2.3 Children's Reading Conversational Agents (CAs)

Reading CAs have moved from screen-based, text-driven helpers to voice-enabled partners that approximate human guidance in vocabulary, comprehension, and motivation [15, 37, 44, 65]. This evolution leverages speech recognition, natural language processing, and dialogue management to implement turn-taking, contingent prompts, and narrative alignment with children's activity.

Representative systems have demonstrated success in specific areas. *Reading Partner* uses scaffolded dialogue with preschoolers and achieves on-topic response rates above 85 percent [66]. *Elinor* integrates conversational prompts into educational programming and improves content comprehension relative to non-interactive baselines [63]. *Rosita Reads* embeds a bilingual agent in e-books to support Latinx families' shared reading and reveals distinct child–parent–agent interaction patterns [62]. *StoryBuddy* flexibly incorporates parental involvement through co-reading and configurable bot-reading [74]. *TaleMate* facilitates parent–child joint reading with voiced, character-based interaction [52]. Beyond literacy narrowly defined, *StoryCoder* shows how interactive storytelling can introduce computational ideas for ages five to eight while cultivating narrative comprehension and creativity [12]. Collectively, these studies show that CAs can increase narrative-relevant talk and achieve comprehension gains comparable to human partners[59, 64].

Despite these advances, a closer analysis reveals that reading CAs still suffers from the fundamental limitations discussed previously. First, they remain predominantly voice-dominant, lacking the multi-modal sensing for non-verbal cues like gaze and gesture that define effective human joint reading. While few studies, such as Minnie, indicate that responsive gaze and facial display can deepen engagement, such capabilities are not yet common in reviewed works[33, 66, 70]. Second, their emotional intelligence is shallow, with limited adaptation to boredom, frustration, or excitement [6, 55]. Third, personalization rarely extends beyond performance-based difficulty adjustments, failing to model individual differences in attention, interaction style, or learning needs[2, 9, 24]. These issues are amplified in real-world settings, where children's open-ended questions meet brittle dialogue policies and commercial products expose the limits of constrained content libraries [1, 27, 46, 60]. Commercial products such as Codi [25], Luca [28], and Luka [29] illustrate the appeal of turnkey storytelling, yet they also expose the limits of manually authored prompts and constrained content libraries.

These persistent gaps strongly motivate the need for multi-modal, affect-aware, and parent-inclusive CAs that synchronize visual cues with dialogue and adapt to individual learning profiles—the directions advanced by the present work.

3 Formative Study

To design and implement an improved AI agent for joint reading with children to enhance reading engagement and efficiency, we conducted a formative study with parents to (1) understand challenges in children's reading engagement and joint reading practices, (2) gather parental perspectives on the proposed design considerations, (3) identify the specific design features of the system, and (4) understand parental concerns regarding privacy, technology integration, and educational value. We adopted a parent-centered approach because parents are the primary orchestrators of home reading and can report reliably on engagement patterns, routines, and constraints in situ [60, 66, 75]; children in this age range often struggle to verbalize metacognitive strategies and design rationales in interviews, making parents appropriate proxies at the formative stage; and prior work has successfully grounded child-facing reading systems in formative studies with parents, including StoryBuddy and StoryDrawer [72, 74].

3.1 Method

We recruited 16 families with elementary school children between ages 6-10 as our parent participants (PPs) for formative study. This age range was deliberately chosen based on developmental research suggesting it represents a critical period in reading development when children transition from *"learning to read"* to *"reading to learn"* [43, 50]. This developmental window is characterized by rapid growth in metacognitive abilities and reading strategies, making it an ideal focus for interventions aimed at enhancing reading engagement and comprehension [15, 24].

To ensure diversity of perspectives, our recruitment strategy encompassed two distinct communities in China. The first community primarily comprised university faculty families with advanced academic credentials (master's or doctoral degrees). These families typically operated in bilingual (Chinese-English) environments, and their children attended English-speaking international schools. The second community represented working-class to middle-class families whose children attended standard Chinese-speaking public schools. We present the demographic characteristics of our participants in Table 1.

Table 1. Demographic information of formative study participants.

ID	Parent Gender	Parent Age	Parent Occupation	Child Age	Weekly Reading (hrs)
PP1	F	45–50	Self-Employed	10	5+
PP2	F	40–45	Education	8	7+
PP3	F	35–40	Education	6	5+
PP4	F	35–40	Business Professional	10	3.5–4
PP5	F	40–45	Business Professional	9	4.5–7
PP6	F	40–45	Business Professional	10	4.5
PP7	F	35–40	Business Professional	9	7–10
PP8	F	30–35	Business Professional	7	3.5–4
PP9	F	40–45	Business Professional	7 & 9	3.5–5
PP10	M	35–40	Tech Professional	6	2–3
PP11	F	35–40	Education	7 & 10	3.5–7
PP12	F	35–40	Education	9	3.5–4
PP13	F	35–40	Self-Employed	8	3.5–4
PP14	F	30–35	Consultant	8	4–6
PP15	F	30–35	Business Professional	6	3.5–4
PP16	F	35–40	Business Professional	8	2.5

Each 60-minute session with a parent began with consent and a brief overview of the study aims. We then explore each family's existing practices through semi-structured interviews. Questions addressed parental observation of children's reading behaviors during joint reading and independent reading, challenges in reading engagement, the integration of technology in educational activities, and previous experience with conversational agents. This approach, aligned with contextual inquiry methods [21], allowed us to understand the lived experiences and specific needs of each family. After surfacing challenges, we facilitated a parent-centered contextual design session to generate design considerations addressing those challenges. Parents are encouraged to articulate desired features, interaction flows, and

even concerns about technologies. Informed by co-design methods adapted for individual engagement [38, 47], this flexible contextual design session produced a set of design considerations and preliminary requirements that informed subsequent system design.

All sessions were audio-recorded and transcribed, with field notes capturing notable reactions and design ideas. We conducted reflexive thematic analysis [3] using an inductive and data-driven approach. Two researchers independently open-coded transcripts and notes in Atlas.ti, memoing emergent concepts and relationships. After coding an initial subset, we met to merge codes into a shared, evolving codebook; we then applied it to the remaining data, revisiting earlier segments via constant comparison until no new themes emerged. Differences were resolved by negotiated agreement. We also synthesized the prioritization activity to surface high-value features and common barriers to adoption.

3.2 Key Insights

Across sixteen formative interviews with parents of elementary-aged children, our analysis surfaced seven interrelated key insights (KIs) about how an AI reading companion should support engagement, comprehension, and family practices.

3.2.1 KI1: Engagement is anchored in children's choices and interest fit; task framing undermines focus. Parents consistently described sustained attention when children selected materials themselves and pursued topics of genuine interest. One parent noted that a child reads *"very focused as long as he chooses it himself"* (PP2), while another reflected that *"when this becomes a task for him… that's when distraction happens"* (PP2). Interest misfit quickly eroded attention: *"if the content of the book is not particularly appealing to him, he won't be particularly focused"* (PP1). By contrast, self-chosen goals fueled completion drive: *"once he starts reading a book, he has this obsession to finish it"* (PP2). Families also observed that children naturally share what excites them (e.g., mealtime storytelling and reflections), reinforcing the role of intrinsic motivation in family literacy practices (PP7; PP14). These patterns validate the challenge of maintaining engagement when reading is positioned as an assignment and underscore the value of personalization that prioritizes children's choices and interests.

3.2.2 KI2: Visual elements function best as just-in-time scaffolds; for advanced readers they can distract or dilute. Visual supports—comics, scene sketches, or focused illustrations—often catalyzed both interest and comprehension for complex content: *"when reading history books, he likes those with comics… when he gets into the story plot, he becomes particularly interested"* (PP1). Parents emphasized that a *"visual block… can help children use a more direct or easier-to-understand way to further understand content or immerse themselves in the scenario"* (PP2). However, more proficient readers found visuals unnecessary or even distracting: *"if you want to turn things from the book into pictures for him, I don't think it's necessary… he'll definitely look at the pictures instead"* (PP3). Parents also cautioned against over-humorizing serious texts: *"making serious content humorous… would damage important meanings of the original text"* (PP1). These observations validate the need to tailor visual support to the reader's level and purpose (e.g., vocabulary, abstract scenes), and they highlight specific features such as adjustable visual intensity and a minimal-visual mode.

3.2.3 KI3: Affect sensing should be voice — only, processed locally, strictly opt-in — and some families prefer human support. Parents broadly accepted prosody-based sensing – *"voice recognition is acceptable"* (PP4), but rejected camera-based monitoring: *"I don't like having a camera constantly watching him… I don't like this from the very root"* (PP3). Trust hinged on locality and transparency: *"(Local processing) would make me feel more at ease"* (PP1). At the same time, responsive feedback mattered for motivation: *"if a child gets lots of positive feedback… it will make them… want to*

continue using it" (PP2). Parents also endorsed combining prosody with content signals: *"judging through the child's voice, tone, and... their answer content"* (PP12). Some families expressed a preference for parental accompaniment in place of algorithmic emotion features, validating the importance of clear opt-in controls and offering a specific feature direction: voice-only detection, on-device processing, and transparent participation prompts.

3.2.4 KI4: Parents want lightweight, objective oversight and discussion highlights, not test-like correctness. Rather than granular grading, parents valued succinct, objective traces that make learning legible: *"emotional things are not as good as objective things—directly seeing and feeling the changes"* (PP4). They asked for synthesized conversation summaries and topic highlights: *"we can generate discussion highlights... letting parents judge if the child understood the content"* (PP5); they were also curious about how the child reasons with AI because *"this process actually exercises children's thinking"* (PP2). Conversely, school-like checks felt counterproductive: *"if there are standard right and wrong answers, it feels a bit like testing"* (PP3). Some appreciated automated progress indicators for convenience (e.g., simple proficiency bands easing supervision) (PP13), but the prevailing preference was for reflective visibility instead of correctness drills. These views both validate frustrations with exam-centric framing and point to specific dashboard features such as time-on-task, themes pursued, and two-to-three talking points for follow-up at home.

3.2.5 KI5: Personalization should adapt to ability, attention, language strengths, and fatigue with micro-pacing and targeted feedback. Parents described tangible benefits from pacing and segmentation: *"through segmented, phased reading, he becomes more interested"* (PP11), and recommended restorative breaks when attention wanes: *"let them rest—play some gentle music and suggest they stretch or close their eyes"* (PP2). Tailored encouragement supported uneven skills, especially in weaker languages (e.g., *"you expressed very clearly just now"*) (PP2). Developmental calibration was equally important: *"for higher grades, you need to add more content about plot questioning and moral aspects"* (PP12). Intra-family variability further underscored the need for graduated support: twins of the same age required different scaffolds and expectations (PP15). These accounts validate attention and ability as dynamic constraints in home reading and motivate features such as micro-sessions, fatigue-aware prompts, grade-aware questioning, and strength-based feedback.

3.2.6 KI6: Safety and governance hinge on parent pre-testing, child-specific filtering, and layered guardrails. Parents expressed a desire to *"try it myself first, and if there are no problems, then let the child use it"* (PP1), shaped by prior negative experiences with *"unhealthy"* recommendations (PP1). Comfort increased with child-specialized safeguards and content review; otherwise, *"concerns about AI outputting uncontrolled content"* persisted (PP16). Families repeatedly endorsed layered oversight, such as multiple agents supervising content and terminating problematic conversations when needed (PP11; PP15), and they reported being *"more willing to let him try"* when outputs were verified or filtered for children (PP5). These views validate safety as a first-order adoption barrier and identify concrete feature directions.

3.2.7 KI7: The parental role shifts from co-reader to coach; tools should enable follow-up, not takeover. Developmentally, families move from joint reading to independent practice: *"then suddenly he read all books by himself"* (PP3), with many households sustaining a shared atmosphere where *"I read my book and he reads his"* (PP1). Parents wished to remain informed without replacing the child's agency: *"we want to know roughly where he struggles rather than having AI... solve problems"* (PP3). They also highlighted the value of succinct context to sustain conversation: without shared traces, *"after hearing him I can only nod or express agreement"* (PP7). The request for post hoc highlights recurred as a preferred mechanism for home dialogue (PP5). These reports validate evolving joint-reading practices and identify specific features that keep parents meaningfully involved, including succinct highlights, misconception cues, and suggested prompts, while preserving children's ownership.

3.3 Design Considerations

Synthesizing the prior work and our findings from the formative study, we determined four design considerations (DCs) for designing and implementing an enhanced multi-modal AI agent for joint reading with elementary school children:

DC1 **Emotion detection from voice and text.** Parents endorsed voice-only, locally processed emotional detection without camera (KI3), and asked that signals be used for moment-to-moment support rather than test-like evaluation (KI4). Therefore, the system reads children's attention and emotion from prosody and language signals and processes these signals on the device to protect privacy. These signals adjust pacing and encouragement and are never used for scoring.

DC2 **Supportive visual cues for comprehension and engagement.** Parents described visuals as effective scaffolds but potentially distracting for advanced readers (KI2). The system creates context-specific illustrations that clarify abstract scenes and help sustain interest. The illustration will be invasive during the reading to avoid distraction.

DC3 **Three-step dialogue for deeper thinking.** The agent guides a three-step dialogue: understand, reason, connect. It adapts difficulty, timing, and feedback to the child's ability, attention, and fatigue. This structure preserves autonomy and interest by letting children pick topics (KI1), while graduated questioning align with parents' calls for developmental calibration (KI5). The system also utilizes multiple agents to supervise the conversation, addressing parents' concerns about content safety (KI6).

DC4 **Parent dashboard for reflective follow-up.** Parents asked for succinct highlights to facilitate mentorship and communication (KI7). The dashboard surfaces themes, key questions, and moments of struggle or high interest to support meaningful conversations at home. It favors simple progress signals and avoids test-style scoring.

4 PolyTaler

4.1 System Overview

Following these four design considerations, we developed PolyTaler, an AI-enabled interactive system that enhances children's reading engagement through a multi-modal approach. PolyTaler incorporates features that allow the agent to collect multi-modal signals from children during the reading and interaction process, and to adaptively generate multi-modal demonstrations in response to their feedback. In this way, the system fosters more dynamic, engaging, and personalized reading experiences for young readers.

The front-end interface was implemented using Streamlit, a Python framework for interactive, data-driven web applications. Its web-based architecture enables PolyTaler to run directly from standard browsers across any device with a speaker and a microphone. As illustrated in Figure 1, PolyTaler consists of four modules, each addressing one design consideration: The visual cues generation module (A) generates context-specific illustrations under quick processing and content safety supervision, thus sustaining engagement while minimizing distraction (DC2). The emotion detection module (B) processes acoustic and language signals locally to infer children's attention and affect, enabling adaptive pacing and encouragement without using cameras (DC1). The conversational agents module (C) guides a three-step dialogue—understand, reason, connect—adapting difficulty and timing while ensuring content safety through multiple-agent supervision (DC3). Finally, the parental dashboard (D) provides succinct highlights of themes, key questions, and moments of struggle or high interest to support reflective conversations at home, while avoiding test-style scoring (DC4).

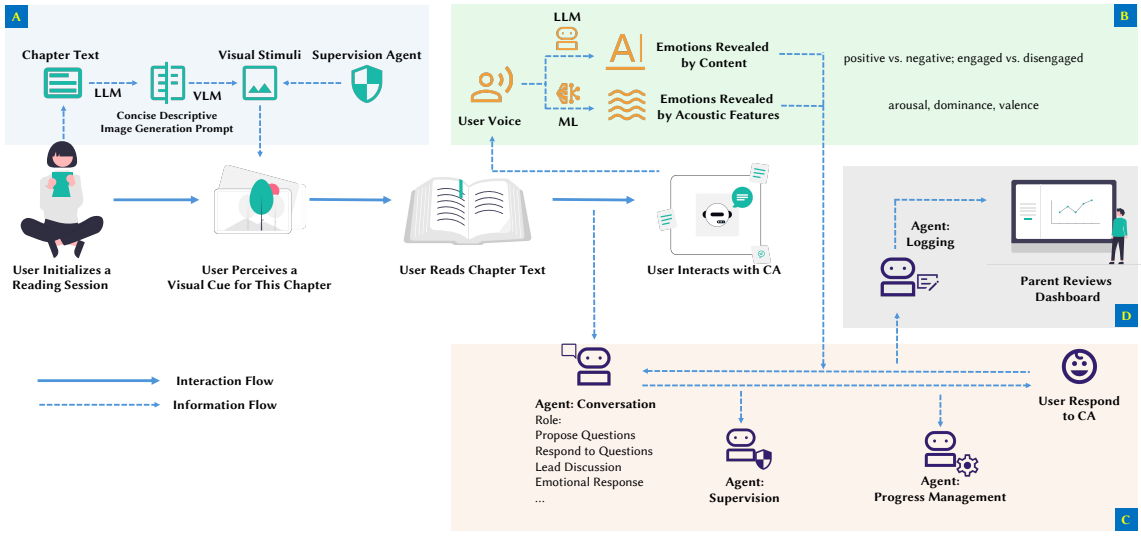

Fig. 1. The System Architecture of PolyTaler: The solid line represents the interaction flow, from users initializing a reading session to interacting with CA; The dashed line represents the information flow, where data are processed through (A) visual cues generation module, (B) emotional detection module, (C) conversational agents module, and (D) parental dashboard.

4.2 Interface and Interaction flow

Following the design considerations, we define the interaction flow as follows. A child initiates an interactive reading session by selecting a book from the welcome interface. At the beginning of each new chapter, the system generates a relevant illustration image that provides visual context and activates schema knowledge prior to reading (Figure 2A, Section 4.3, DC 2). After engaging with these pre-reading visual cues, the child transitions to the reading page and proceeds to read the chapter content in an e-book format (Figure 2B). Upon completing a chapter, the child can click a button to initiate a conversation with the AI reading companion. The conversational agent appears persistently on the right-hand side of the reading page in a chatbot format, presenting an initial question and then communicating with a natural-sounding, emotionally expressive voice and then activating a microphone channel for verbal responses of child (Figure 2C). Concurrently, the agent monitors and interprets the child's emotional states through multi-modal audio–text analysis (Section 4.4, DC1). By combining the child's textual responses with detected affective cues, the system adaptively generates contextually relevant follow-up questions that foster critical thinking and deliver personalized feedback (Figure 2D). When incorrect responses are detected, the conversational agent integrates both textual and emotional signals to generate scaffolding hints that guide the child to retry. The child then continues to engage in subsequent rounds of questioning. The whole process was guided by a three-step dialogue strategy to trigger the child's deeper thinking (Section 4.5, DC3). Finally, all interactions are systematically recorded by a logging agent and presented through a parent-facing dashboard (Figure 2E, Section 4.6, DC4), enabling caregivers to monitor both reading progress and emotional engagement while preserving the child's autonomy during the reading process.

4.3 Visual Cues Generation

Visual cues support children's comprehension and sustain engagement. As shwon by Figure 1A, at the beginning of each chapter, the system receives the chapter content and uses a pre-trained LLM to compose a concise prompt

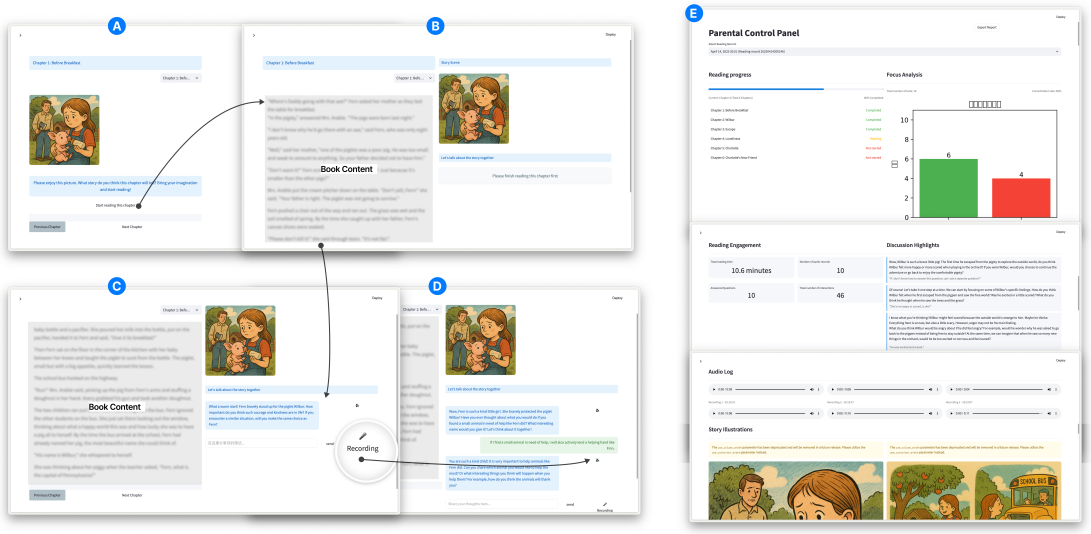

Fig. 2. Reading Page UI: (A) After finishing the visual cues, the user click on "Start to Read"; (B) The user reads the chapter text without distraction; (C) When the user finishes the chapter, he clicks on "Finish Reading" to start to talk with the conversational agent; (D) The conversational agent will generate questions and encourage the user to interact. User communicates with the conversational agent through voice, enabling real-time emotion detection via dual-pathway analysis; (E) After finishing reading, parents can view detailed data and highlights from parental dashboard.

that captures key characters, settings, events, and causal relations. This prompt is then passed to GPT-4o to generate illustration-style visual cues. A pre-trained supervision agent reviews each image for fidelity to the story, adherence to style requirements, and suitability for children; images that fail any check are rejected and regenerated. The approved visuals serve as pre-reading cues that activate prior knowledge, aid learners who benefit from pictorial representations, and build anticipation for the narrative that follows.

4.4 Emotion Detection System

PolyTaler employs a multi-modal framework for emotion detection, integrating both textual and audio signals to construct a comprehensive understanding of children's affective states during interaction, while preserving children's privacy. This framework combines LLM for textual emotion recognition with a machine learning model for vocal affect analysis, as shown by Figure 1B. A fine-tuned LLM is used to analyze children's utterances and classify affect along two axes: positivity versus negativity and engagement versus disengagement. In parallel, the audio module fine-tunes a Wav2Vec2 transformer for vocal affect. This model assesses emotional states across three widely recognized dimensions of affect: arousal (vocal energy and intensity), dominance (confidence or perceived control), and valence (positivity versus negativity of expression).

The system then passes both language and acoustic emotion signals to the conversation agent to guide moment-to-moment support. These emotional signals will affect the agent's response generation, allowing the system to adjust pacing, encouragement, and question difficulty to match the child's current state. These signals are received by the conversational agents.

4.5 Personalized Conversational Agent

The conversational agent module generates context-appropriate prompts and responses tailored to each child's reading behavior, affect, and engagement. It consists of a main agent handling conversation with a supervision agent ensuring content safety, a progress management agent coordinating conversation stages, and a logging agent recording structured traces, as shown by Figure 1C. The module maintains a turn-level state that includes recent answers, explanation quality, attention and affect estimates, and reading pace. This state anchors the dialogue within a structured three-stage educational framework and regulates the LLM through carefully designed prompt engineering, ensuring that every response remains consistently aligned with the intended educational objectives.

4.5.1 Emotionally Responsive Interaction. Multi-modal signals from the emotion detection system directly shape the dialogue strategy. Emotional cues are injected into the LLM's input context and control variables such as target reading level, hint frequency, and wait time before follow-up prompts. When the system detects disengagement, frustration, or confusion, the agent shortens turns, simplifies language, adds step-by-step scaffolds, and introduces supportive feedback. When it observes curiosity or sustained positive affect, the agent lengthens turns, raises question difficulty, and invites deeper reasoning.

4.5.2 Three-Stage Conversation Framework. We combine the Dialogic Reading methodology [71] with the Metacognitive Strategy Development [43] to form an educational framework. This framework guides the conversational structure, which follows three adaptive stages: comprehension, reasoning, and personal connection. The first stage checks factual understanding and basic narrative structure. The second stage elicits predictions, inferences, and text-to-self or text-to-world links. The final stage encourages reflection on relevance and emotion. Progression is gated by lightweight criteria derived from the turn-level state, such as correctness over recent turns, the presence of evidence in answers, and stable attention. If comprehension remains weak, the agent cycles within earlier stages until the criteria are met.

4.5.3 Multi-Agent Personalization Architecture. Personalization is coordinated by three agents with distinct roles. The main agent produces child-facing dialogue conditioned on the story context and the current state. It is prompted to propose questions, respond to questions, lead discussions, and provide emotional support. The supervision agent monitors fidelity to literacy goals, detects topic drift, and most importantly, ensures the content is safe for children. The progress management agent tracks engagement and comprehension trajectories and decides when to conclude a segment or advance the content. A structured prompt-routing mechanism passes summaries of the ongoing exchange and the state to these agents in turn, merges their recommendations, and issues the final prompt to the LLM. This design preserves conversational continuity while enabling oversight and incremental adaptation.

4.6 Parental Dashboard

The parental dashboard serves as an interactive tool that enables caregivers to monitor and interpret their child's reading engagement. As shown in Figure 2E, it presents progress tracking, focus analysis, engagement metrics, conversational highlights, and session-level multimedia logs, supporting informed parental involvement while maintaining the child's autonomy.

A curated highlight section surfaces meaningful exchanges such as emotional reflections or reasoning steps, offering insight into the child's comprehension and thought process. Parents can also review session logs, including timestamped audio clips and linked visual story excerpts that connect spoken responses to narrative content. The dashboard design

Table 2. Demographics of family participant pairs (FP1–FP12). Reading level codes: High = Higher than average, Avg = On average, Low = Lower than average.

FP	Parent's Gender	Parent's Age	Child's Gender	Child's Age	Self-Reported Reading Level	Weekly hrs
FP1	F	40–45	M	9	High	5–7
FP2	F	35–40	M	9	Avg	7–10
FP3	F	35–40	M	10	High	3–4
FP4	F	40–45	F	10	Low	2–3
FP5	F	40–45	M	7	High	2–3
FP6	F	45–50	M	8	Avg	3–4
FP7	F	45–50	F	7	Avg	3–4
FP8	F	40–45	M	6	High	3–4
FP9	F	40–45	M	8	Low	1–2
FP10	F	35–40	F	10	Avg	3–4
FP11	M	35–40	F	8	Avg	3–4
FP12	F	45–50	F	10	High	7–10

emphasizes clarity and interpretability, allowing parents to easily identify trends and emerging patterns without requiring technical expertise.

5 User Study

5.1 Participants

In order to understand how children perform with the assistance of PolyTaler and how parents perceive its value and effectiveness, we recruited 12 children accompanied by their parents as family participants (FPs) through the snowball sampling method [39] for the user evaluation. The family participant pairs (FP1–FP12) consisted of 10 pairs whose parents had previously participated in our formative study, along with 2 additional pairs (FP6 and FP9), as shown in Table 2. All participants resided in China and were fluent in Chinese. Each pair received compensation of CNY 100 for their time and effort.

5.2 Procedure

We invited the child participants, accompanied by their parents, to the university's lab to conduct user studies (Figure 3). At the beginning of each session, participants received a standardized introduction outlining the procedure and expectations. Throughout the experiment, we recorded screen activities and audio, deliberately omitting video recording of participants' faces to address privacy concerns. Parents remained present during the sessions and kept quietly observing the interaction between child and the system while positioned behind the child, out of the child's direct view to ensure that their presence did not interfere with the interaction.

Although the system can ingest any e-book, we standardized on *Charlotte's Web* to maintain internal validity and cross-participant comparability. This book is an age-appropriate middle-grade text with moderate difficulty and clear chapter boundaries [41]. To control for potential order effects and content familiarity, we implemented a counterbalanced design. For the first 6 child participants (FP1-FP6), they read the first two chapters of *Charlotte's Web* using our experimental system, followed by reading chapters 3-4 in the control condition without the system. For the remaining 6 participants

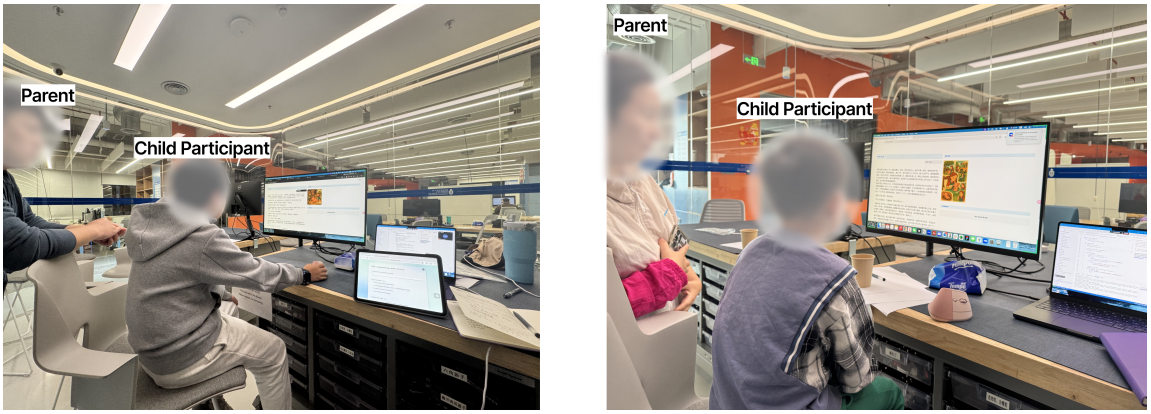

Fig. 3. User study sessions set up. The images are taken from two different sessions. The images are partially blurred to protect the privacy of the participants.

(FP7-FP12), the sequence was reversed, where they began with chapters 1-2 in the control condition before experiencing chapters 3-4 with the system. This approach helped minimize biases that might arise from order effects or differences in chapter content difficulty.

After each reading session, children completed a comprehension assessment and an engagement questionnaire. The comprehension assessment included multiple-choice questions designed to evaluate various aspects of reading comprehension, including main idea identification, character understanding, causal relationships, and inferential reasoning [40]. The engagement questionnaire used a 5-point pictorial Likert scale to assess interest, attention, and motivation to continue reading [16]. For the experimental condition, additional questions evaluated the perceived helpfulness, usability, and comprehensibility of the AI assistant [66].

At the end of each reading session, parents were invited to interact with the parental dashboard and engage in a brief discussion with their children to conclude the activity. Following this, we conducted semi-structured interviews, beginning with children and then with parents separately, to capture individual perspectives without mutual influence. Children were asked about their preferences between reading methods, perceived comprehension support, feature preferences, and suggestions for improvement. Parent interviews focused on observed differences in their child's engagement, perceived benefits of the AI assistant, comparisons with traditional parent-child reading, concerns about the technology, and potential home use scenarios. These interviews typically lasted 15-20 minutes and provided valuable qualitative insights complementing our quantitative measures. The procedure for participants is shown in Figure 4.

We analyzed the data using a mixed-methods design that combined quantitative outcomes with qualitative evidence. On the quantitative side, we compared post-session multiple-choice comprehension scores and 5-point pictorial engagement ratings between the AI-agent and control conditions using paired-samples t-tests ($\alpha = .05$; $n = 12$), checked normality of paired differences with Shapiro-Wilk tests, and summarized system interaction logs with descriptive statistics. All analyses were run in Python. On the qualitative side, we conducted reflexive thematic analysis [3] with an inductive, data-driven approach. Two researchers independently open-coded transcripts and notes in *Atlas.ti* and wrote brief memos on emerging ideas. After coding an initial subset, we met to merge codes into a shared, evolving codebook, applied it to the remaining data, and revisited earlier segments through iterative comparison until no new themes appeared; disagreements were resolved through discussion and negotiated agreement.

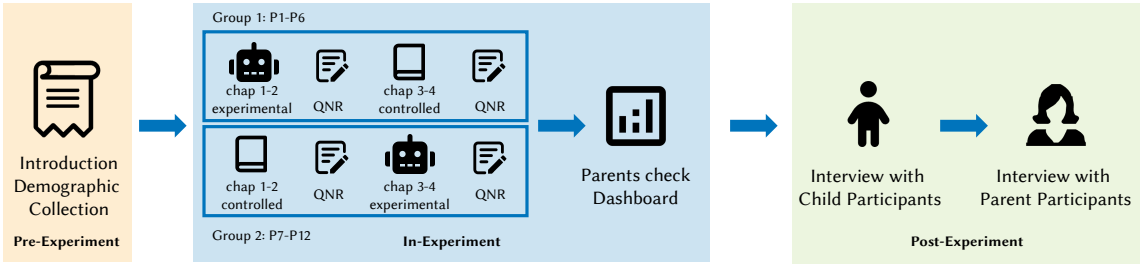

Fig. 4. The procedure for participants

5.3 Findings

All twelve children completed two reading sessions: one with the prototype and one without. The average reading time was 5 minutes 6 seconds with the prototype and 7 minutes 35 seconds without it. Following the sessions, each child completed the comprehension assessments and participated in post-experiment interviews. All parents also took part in the interviews. On average, each experiment session lasted 58 minutes.

5.3.1 How Did Children Interact with PolyTaler? Children's interactions with PolyTaler typically followed our three-step dialogue design, understand → reason → connect, while children actively steered pacing and topic. Sessions often opened with values-forward comprehension prompts that invited personal stance-taking rather than simple recall. For example, the agent asked, *"If you met a small animal that needed help, would you stand up like Fern, and why is it important to protect the weak?"* (Q, PolyTaler). A child replied, *"I would stand up like Fern. Killing a pig just because it is small and weak feels wrong. We should let it grow up, and I would help the animal"* (FP1, Child). Children also demonstrated accurate story grounding, for instance, *"I think Charlotte will spin words in her web, like 'Some Pig,' to save Wilbur"* (FP1, Child). These initial exchanges showed affective buy-in and a quick bridge from events in the chapter to moral reasoning.

Pre-reading visual cues, which primed characters and settings, appeared to ease entry into the conversation. When the agent followed up with, *"Who do you think the gentle voice at the end might be, and why?"* (Q, PolyTaler), children immediately drew on the story world: *"I think it is the spider"* (FP4, Child) and *"It might be Fern"* (FP6, Child). Visual context also supported anticipatory reasoning. After a brief scene recap, one child, asked to reflect on Wilbur's escape, predicted, *"He felt excited and scared"* (FP7, Child). These examples suggest that children used the illustrations as anchors, making it easier to name entities, keep track of events, and engage in causal discussion without an extended warm-up.

When questions were initially too abstract, the agent recalibrated by using stepwise scaffolding and emotion-aware pacing. After one child noted, *"I did not understand the question"* (FP3, Child), the agent reframed it: *"Let us take it step by step. When Wilbur first escaped, did he feel excited or a little scared?"* (Q, PolyTaler). The child then provided a mixed-affect answer and continued the conversation. Children also signaled attention changes directly: *"Do you have more questions for me?"* (FP2, Child), *"Let us change the question"* (FP1, Child), and *"I do not want to talk about this topic"* (FP4, Child). In these moments the agent shortened turns, added encouragement, or pivoted with a context-preserving probe, for example, *"Who might help Wilbur next, and what makes you think so?"* (Q, PolyTaler). When asked to consider safety, one child responded with self-regulation, *"If I heard a loud bang, I would probably run straight back to the pen. It*

is a bit early for me to go out into the world" (FP7, Child). These adjustments exemplify DC1 in practice, where detected boredom or frustration resulted in gentler pacing and supportive tone rather than performance scoring.

Children occasionally moved off topic or engaged in playful boundary-testing, and the agent guided them back while maintaining rapport. A child invited role-play, *"Can you speak as Wilbur? I want to ask fun questions"* (FP4, Child), and later joked, *"It would turn into braised pork"* (FP4, Child). The agent acknowledged the humor and redirected toward constructive planning: *"If Wilbur explores, what should he do to stay safe, and who could help?"* (Q, PolyTaler). This prompted children to produce grounded plans and constraints, for example, *"They would need a rope. Wilbur is small and cannot climb, so his former owner could carry him up. It would be funny"* (FP5, Child). Curiosity about modality also appeared, as when one child asked, *"How can a spider talk with a pig?"* (FP4, Child), which the agent answered briefly before returning to the chapter's goal.

By the connect stage, children tied the story to everyday action and personal preferences. Following the prompt, *"In your own life, what could you do to protect small lives?"* (Q, PolyTaler), one child committed, *"Not throwing trash into the ocean can protect sea life"* (FP2, Child). Preference-based talk was specific and social, for example, *"I want to meet a cat and play with a yarn ball on the grass"* (FP6, Child). Taken together, interactions showed a consistent flow from visual priming and comprehension, to calibrated causal reasoning supported by emotion-aware pacing, and finally to personal connection and concrete plans for action. This flow illustrates DC2 as an effective on-ramp to dialogue and DC1 as a quiet regulator that helped sustain engagement when attention dipped or topics felt difficult, while DC3 serves as an overall structure of the interaction.

5.3.2 How Did Children and Parents Perceived PolyTaler? Children approved and enjoyed PolyTaler as evidenced by questionnaires, behavioral logs, and semi-structured interviews. Parents, after observing the sessions and engaging with the dashboard, were also supportive of the system. Both groups offered targeted suggestions for further refinement. We report seven Key Findings (KFs) that integrate convergent and divergent evidence across data sources.

KF1: PolyTaler reliably elevated interest and attention. Paired-samples tests show large, statistically significant improvements in self-reported reading interest ($d \approx 1.34$, p=.0063) and attention/focus ($d \approx 1.45$, p=.0061) with the agent, as shown in Figure 5. Parents reported observing stronger children's engagement during sessions. Parents noted that *"Her participation improved a lot (from my observation)"* (FP10, Parent), while another remarked that *"the participation level is definitely much higher; when reading books normally (without the system), I even don't know what he's thinking"* (FP1, Parent). Children echoed this experiential lift as well, stating that follow-up turns kept the exchange *"interesting"* (FP7, Child) and that encouragement made chatting feel *"relaxed"* (FP6, Child). Taken together, the quantitative and qualitative evidence indicate that the agent reliably increases affective engagement and on-task behavior.

KF2: Questioning scaffolds deeper thinking, but must be appropriately challenging. In interviews, participants credited the agent's questioning and explanations with nudging comprehension, recall, and connection-making. This finding is supported by high feature ratings (e.g., helpfulness of questions, $M \approx 4.25$; explanations for understanding, $M \approx 4.33$; understanding/response ability, $M \approx 4.83$), as shown in Figure 6. Children reported that the PolyTaler *"helps me understand … he will ask some questions for me to answer"* (FP8, Child) and that the *"Q&A is most helpful for improving reading ability"* (FP10, Child). At the same time, several participants emphasized the need for calibration: some questions felt *"a bit difficult … too broad"* to answer immediately (FP2, Child), and one parent noted the opening sequence could *"feel like doing reading comprehension,"* which dampened enjoyment (FP4, Parent). Another child described certain prompts as *"a bit brain-twisting"* (FP12, Child). Quantitatively, comprehension scores increased on average (+19.4%, $d \approx 0.70$) but were

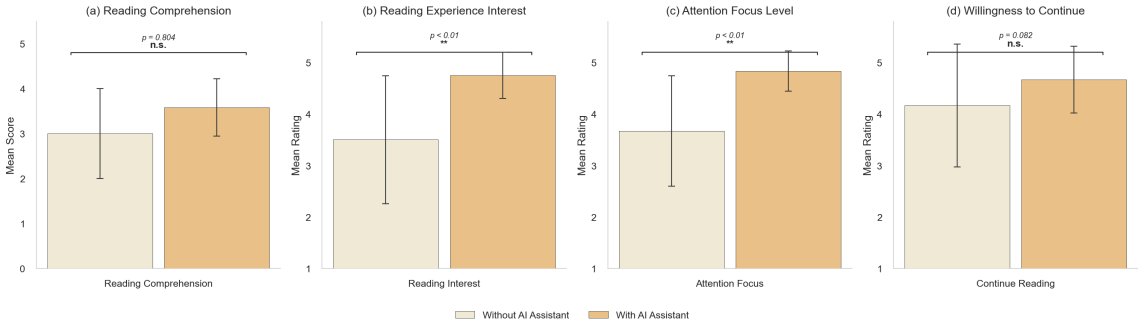

Fig. 5. Reading comprehension and engagement scores with vs. without PolyTaler.

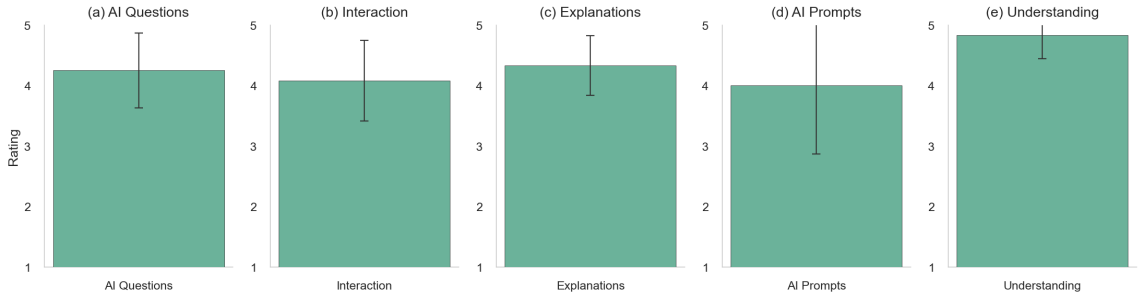

Fig. 6. Features Evaluation: Helpfulness of AI Questions, Ease of AI Interaction, AI Explanations, Helpfulness of AI Prompts, and AI Understanding and Response Ability

not statistically significant in this sample, a pattern consistent with reports that benefits depend on difficulty, timing, and child preference. Taken together, these results suggest that questioning can strengthen children's understanding and help them connect ideas, such as recalling details, explaining characters, or linking events to their own experiences, when the questions are pitched at an age-appropriate and manageable level.

KF3: Supportive visuals aid comprehension and stamina, but should be scheduled before and after chapters, not mid-sentence. Children frequently attributed better understanding to context-specific illustrations (e.g., *"because it has those illustrations, let me understand better"* (FP2, Child); *"I prefer it because there are pictures and voice"* (FP3, Child)). Parents highlighted their value for children transitioning from picture-heavy to text-heavy books (*"pictures help him get a general idea of the story"* (FP4, Parent)). However, participants asked us to avoid intrusive mid-sentence pop-ups. One child stated plainly: *"better after reading . . . reading halfway and interrupting makes the impression not deep"* (FP3, Child), and a parent worried that dynamic visuals could distract from the reading flow (FP5, Parent). This feedback points to a clear visual policy: present a brief character/setting primer *before* reading; offer a concise recap board *after* each chapter; and allow the child to press a clear button to see more or fewer images on demand. This schedule preserves focus while retaining the benefits of visual scaffolding.

KF4: Voice and persona of CA shape user's comfort and expressivity. Voice-based interaction helped children organize and express their thoughts. As one parent observed, *"through voice AI communication, you can help him summarize what*

he wants to say" (FP1, Parent). Both children and parents requested warmer, more playful, or story-character voices: several found the default rendering *"too harsh"* or *"a bit stiff"* and preferred *"a boy's voice or a child's voice of similar age"* (FP6, Child) or a more *"cartoonish"* style (FP12, Child). A child suggested that *"characters from the story . . . ask questions in their roles"* (FP3, Child). Parents also remarked that AI feels like an *"equal"* partner compared with parental authority, lowering performance anxiety and eliciting freer responses (*"he can say anything to AI without emotional feedback"* (FP10, Parent)). These accounts support adding persona packs and prosody controls that match age and context, and occasionally embody in-story characters to sustain rapport without sacrificing guidance.

KF5: Personalization that prioritizes children's choices improves acceptance and depth. Families requested personalization options that let children steer the experience. One child wanted the *"number of questions"* to be balanced with story content (FP2, Child), while a parent asked for *"conversations suitable for children of the same age"* (FP4, Parent). Another child wanted interactions to be *"more humorous, more fun"* when energy wanes (FP10, Child). Parents emphasized developmental variation — *"different age groups need different designs"* (FP9, Parent) — and situational factors such as fatigue and bedtime routines. This pattern, combined with high ratings for interaction ease and understanding, motivates a child-facing control strip for immediate adjustments (e.g., more/less questions, lighter/deeper exploration, more/fewer pictures, preferred voice/persona). Such a feature would reduce any *"homework vibes"* while preserving cognitive challenge.

KF6: The parent dashboard enables reflective follow-up, but needs sharper conversation starters. Parents valued transparent records of reading activity and safety: *"we can see . . . reading progress . . . attention analysis . . . all conversation history . . . plus all AI-generated illustrations"* (FP10, Parent). After sessions, they used focus summaries (*"six times focused and four times unfocused"* (FP12, Parent)) and discussion logs to quickly gauge the child's experience. However, they also wanted more concise, actionable prompts: *"at least I know what he read"* (FP7, Parent), but they also asked for short summaries and indicators of understanding (*"how many did he get right . . . does he really understand this book?"* (FP7, Parent)) and even *"mind-map keywords"* to seed dialogue (FP7, Parent). Parents cautioned against test-style grading. This feedback points to a direct design implication: provide a one-minute *"Talk Together"* card that shows three key ideas, two short child quotes, and one or two seed questions—clear signals without scores.

KF7: Boundaries and context require avoiding dependency, honoring preferences, and mitigating screen/UX costs. Parents raised concerns about potential *"reliance"* on the tool (FP1, Parent) and reduced social practice (FP2, Parent). Several noted eye strain and lower immersion with screens (*"your eyes can't handle it" and "immersion not as good as books"* (FP11, Parent)). Some children preferred *"silent reading"* (FP6, Child) or independent reading to avoid pressure (*"I prefer reading by myself"* (FP4, Child)), underscoring that AI should complement, not replace, human-led reading. Parents nevertheless saw practical value when adults are unavailable (*"replace some time when adults are not accompanying"* (FP4, Parent)) and framed AI as *"complementary"* to books (FP11, Parent). Practical constraints—such as responsiveness and smooth scrolling—also mattered (FP5, Parent). Accordingly, the system should be positioned explicitly as assistive; include session timers and break nudges; offer an eye-comfort display mode (e.g., low-glow/E-ink friendly layouts); and ensure low-latency, predictable navigation.

6 Discussion

Our study indicates that a conversational multi-modal companion can enrich at-home reading by raising children's engagement, supporting deeper meaning making, and enabling brief, high-quality parent–child follow up. Children

expressed positive views of pre-chapter visual cues when these were scheduled to avoid interruption of reading. These outcomes address recurring limits in prior reading CAs, where systems are often voice-centered with constrained affect modeling and limited personalization to heterogeneous skills and habits [59, 60, 74]. Compared with CA designs that emphasize voice-only input and pre-authored prompt sets [1, 27], PolyTaler combines voice–text affect signals processed locally, scheduled visual scaffolds, a calibrated questioning framework, and a parent-facing view that surfaces traces for conversation.

Multi-modality in our design is expressed through children's experience and the system's processing: children engage with visual, textual, and audio cues through AI-generated illustrations, the book content itself, and oral conversation with the agent, while the system senses attention and affect from prosody and language signals to adjust pacing and encouragement. In our evaluation, gentle, short early turns built momentum, and when confusion or disengagement appeared, simple rephrasings, two-choice prompts, and brief pauses helped children re-enter the task. Children preferred brief illustrations before and after chapters rather than pop-ups during a sentence, which is consistent with evidence that visuals work best when they prepare or consolidate content at natural pauses [19, 71]. Our questioning moved from understanding to reasoning to personal connection. This flow fits shared-reading practices that begin concrete and then invite inference and links to prior knowledge [56, 71], and it aligns with findings that conversational agents can support comprehension when prompts are contingent and sized to the child's state [63, 64]. Children asked for warm, story-aligned voices and light humour. Prior work supports simple interfaces, approachable character personas, and age-appropriate humour for readers in our age range [31, 49]. These choices address common limits in voice-only designs with narrow affect handling and little room for different skills and habits [59, 61, 72, 74].

These observations refine our design considerations and lead to the following design implications: First, keep affect sensing simple and local. Use only voice and text signals, process them on-device, and limit their role to tuning pacing and encouragement—not grading or profiling. Start with a brief readiness check, keep early turns short, and adapt in the moment when hesitation appears by rephrasing simply, offering a two-choice question, or inserting a short pause. Second, schedule visuals to preserve reading flow: show a 10–20 s primer before a chapter and a 20–30 s recap after, avoid mid-sentence pop-ups, and provide a child-facing toggle for image density along with a minimal-visual mode for advanced readers. Third, include a launch-page control center for question frequency, depth of discussion, image density, and voice, with story-aligned persona defaults that remain warm and safe. Together, these choices share control between child and system, maintain momentum, and personalize the experience without prematurely raising task difficulty or expanding data collection.

Parents in our study valued privacy, clear purpose, and tools that support rather than replace shared reading. We keep sensing to voice and text, process signals locally, and use them only to adjust pacing and encouragement. We do not store profiles and do not grade children, which follows guidance for educational AI that recommends minimal collection, local processing when feasible, and clear explanations of why signals are used [32]. It also agrees with survey work that separates interaction feedback from broad judgments about the child [14, 69]. From parental perspectives, we propose future design implications: First, one-minute "Talk Together" card that lists key ideas, short child quotes, and one or two seed questions gave parents something concrete to say next without turning reading into a test, which echoes work showing that targeted prompts, not scores, help adults support vocabulary and reflective talk. These practices fit how families already use voice agents and interactive media when the experience is simple, brief, and easy to place in a routine [1, 27]. Moreover, for safety, the system should include on-device filtering for sensitive content, a clear "ask a parent" option, and periodic review of prompts and summaries with families. To support access, it should

also run in bandwidth-limited homes and on ordinary devices, with phone-first or low-spec modes, optional offline packs, multilingual support, and local processing.

7 Conclusion

The paper presents PolyTaler, a multimodal AI reading companion designed to support joint reading with elementary school children (ages 6–10), with the goal of enhancing both engagement and efficiency. Guided by our formative study, we distilled four design considerations: (DC1) use on-device emotion detection to adjust pacing and provide encouragement; (DC2) generate clarifying illustrations while minimizing distraction; (DC3) structure a three-step dialogue—from comprehension to reasoning to personal connection—with adaptive difficulty; and (DC4) provide a parent dashboard summarizing themes, key questions, and salient moments for reflection at home.

We evaluated PolyTaler in a user study with 12 child–parent pairs. Findings indicate that PolyTaler effectively leverages multimodal information to sustain children's engagement during storytelling. Both children and parents perceived the system as a trustworthy and engaging partner in joint reading activities.

References

[1] Erin Beneteau, Ashley Boone, Yuxing Wu, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2020. Parenting with Alexa: Exploring the Introduction of Smart Speakers on Family Dynamics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376344

[2] Neelma Bhatti, Timothy L. Stelter, Scott McCrickard, and Aisling Kelliher. 2021. Conversational User Interfaces As Assistive Interlocutors For Young Children's Bilingual Language Acquisition. In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences* (Virtual Event, USA) *(IMX '21)*. Association for Computing Machinery, New York, NY, USA, 208–211. https://doi.org/10.1145/3452918.3465498

[3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[4] Elizabeth Broadbent, Mark Billinghurst, Samantha G. Boardman, and P. Murali Doraiswamy. 2023. Enhancing Social Connectedness with Companion Robots Using AI. *Science Robotics* 8, 80 (July 2023), eadi6347. https://doi.org/10.1126/scirobotics.adi6347

[5] Jennings Bryant and Dolf Zillmann. 2014. Using Humor to Promote Learning in the Classroom. *Humor and children's development* (2014), 49–78.

[6] Jiaju Chen, Minglong Tang, Yuxuan Lu, Bingsheng Yao, Elissa Fan, Xiaojuan Ma, Ying Xu, Dakuo Wang, Yuling Sun, and Liang He. 2025. Characterizing LLM-Empowered Personalized Story Reading and Interaction for Children: Insights From Multi-Stakeholder Perspectives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1002, 24 pages. https://doi.org/10.1145/3706598.3713275

[7] Nian-Shing Chen, Daniel Chia-En Teng, Cheng-Han Lee, and Kinshuk. 2011. Augmenting Paper-Based Reading Activity with Direct Access to Digital Materials and Scaffolded Questioning. *Computers & Education* 57, 2 (Sept. 2011), 1705–1715. https://doi.org/10.1016/j.compedu.2011.03.013

[8] Xiyan Chen. 2020. AI + Education: Self-adaptive Learning Promotes Individualized Educational Revolutionary. In *Proceedings of the 2020 6th International Conference on Education and Training Technologies*. ACM, Macau China, 44–47. https://doi.org/10.1145/3399971.3399984

[9] Carol McDonald Connor, Frederick J. Morrison, Barry Fishman, Elizabeth C. Crowe, Stephanie Al Otaiba, and Christopher Schatschneider. 2013. A Longitudinal Cluster-Randomized Controlled Study on the Accumulating Effects of Individualized Literacy Instruction on Students' Reading From First Through Third Grade. *Psychological Science* 24, 8 (Aug. 2013), 1408–1419. https://doi.org/10.1177/0956797612472204

[10] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems That Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–13. https://doi.org/10.1145/3544548.3580672

[11] Richa S. Deshmukh, Tricia A. Zucker, Sherine R. Tambyraja, Jill M. Pentimonti, Ryan P. Bowles, and Laura M. Justice. 0034/2019. Teachers' Use of Questions during Shared Book Reading: Relations to Child Responses. *Early Childhood Research Quarterly* 49 (0034/2019), 59–68. https://doi.org/10.1016/j.ecresq.2019.05.006

[12] Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L Murnane, and James A. Landay. 2021. StoryCoder: Teaching Computational Thinking Concepts Through Storytelling in a Voice-Guided App for Children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–15. https://doi.org/10.1145/3411764.3445039

[13] Griffin Dietz, Zachary Pease, Brenna McNally, and Elizabeth Foss. [n. d.]. Giggle Gauge: A Self-Report Instrument for Evaluating Children's Engagement with Technology. ([n. d.]).

[14] Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation. https://doi.org/10.48550/arXiv.2305.00955 arXiv:2305.00955 [cs]

[15] Zoe M. Flack, Andy P. Field, and Jessica S. Horst. 2018. The Effects of Shared Storybook Reading on Word Learning: A Meta-Analysis. *Developmental Psychology* 54, 7 (July 2018), 1334–1346. https://doi.org/10.1037/dev0000512

[16] Jennifer A Fredricks, Phyllis C Blumenfeld, and Alison H Paris. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research* 74, 1 (2004), 59–109.

[17] Carme Garcia Yeste, Regina Gairal Casadó, Ariadna Munté Pascual, and Teresa Plaja Viñas. 2018. Dialogic literary gatherings and out-of-home child care: Creation of new meanings through classic literature. *Child & Family Social Work* 23, 1 (2018), 62–70.

[18] Radhika Garg and Subhasree Sengupta. 2020. He Is Just Like Me: A Study of the Long-Term Use of Smart Speakers by Parents and Children. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (March 2020), 1–24. https://doi.org/10.1145/3381002

[19] Lorenz Grolig, Caroline Cohrdes, Simon P. Tiffin-Richards, and Sascha Schroeder. 0032/2020. Narrative Dialogic Reading with Wordless Picture Books: A Cluster-Randomized Intervention Study. *Early Childhood Research Quarterly* 51 (0032/2020), 191–203. https://doi.org/10.1016/j.ecresq.2019.11.002

[20] Kenneth Holstein, Vincent Aleven, and Nikol Rummel. 2020. A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. In *Artificial Intelligence in Education*, Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán (Eds.). Springer International Publishing, Cham, 240–254. https://doi.org/10.1007/978-3-030-52237-7_20

[21] Karen Holtzblatt and Hugh Beyer. 2016. *Contextual design: Design for life.* Morgan Kaufmann.

[22] David P.H Jones. 2002. Editorial: Listening to Children. *Child Abuse & Neglect* 26, 8 (Aug. 2002), 833–835. https://doi.org/10.1016/S0145-2134(02)00352-6

[23] David Comer Kidd and Emanuele Castano. 2013. Reading Literary Fiction Improves Theory of Mind. *Science* 342, 6156 (2013), 377–380. https://doi.org/10.1126/science.1239918 arXiv:https://www.science.org/doi/pdf/10.1126/science.1239918

[24] Irina R. Kumschick, Luna Beck, Michael Eid, Georg Witte, Gisela Klann-Delius, Isabella Heuser, RÃ¼diger Steinlein, and Winfried Menninghaus. 2014. READING and FEELING: The Effects of a Literature-Based Intervention Designed to Increase Emotional Competence in Second and Third Graders. *Frontiers in Psychology* 5 (Dec. 2014). https://doi.org/10.3389/fpsyg.2014.01448

[25] Pillar Learning. 2025. *Codi the Storytelling Robot.*

[26] Arne Lervåg, Charles Hulme, and Monica Melby-Lervåg. 2018. Unpicking the Developmental Relationship Between Oral Language Skills and Reading Comprehension: It's Simple, But Complex. *Child Development* 89, 5 (Sept. 2018), 1821–1838. https://doi.org/10.1111/cdev.12861

[27] Silvia B. Lovato, Anne Marie Piper, and Ellen A. Wartella. 2019. Hey Google, Do Unicorns Exist? Conversational Agents as a Path to Answers to Children's Questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (Boise, ID, USA) *(IDC '19)*. Association for Computing Machinery, New York, NY, USA, 301–313. https://doi.org/10.1145/3311927.3323150

[28] Luca. 2025. *LUCA.ai Reading Platform.*

[29] Luka. 2025. *Luka® the Reading Companion for Kids.*

[30] Larysa V. Lysenko and Philip C. Abrami. 2014. Promoting Reading Comprehension with the Use of Technology. *Computers & Education* 75 (June 2014), 162–172. https://doi.org/10.1016/j.compedu.2014.01.010

[31] Naja A. Mack, Dekita G. Moon Rembert, Robert Cummings, and Juan E. Gilbert. 2019. Co-Designing an Intelligent Conversational History Tutor with Children. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, Boise ID USA, 482–487. https://doi.org/10.1145/3311927.3325336

[32] Fengchun Miao and Wayne Holmes. 2023. *Guidance for Generative AI in Education and Research.*

[33] Joseph E. Michaelis and Bilge Mutlu. 2017. Someone to Read with: Design of and Experiences with an In-Home Learning Companion Robot for Reading. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver Colorado USA, 301–312. https://doi.org/10.1145/3025453.3025499

[34] Suzanne E. Mol, Adriana G. Bus, Maria T. De Jong, and Daisy J. H. Smeets. 2008. Added Value of Dialogic Parent–Child Book Readings: A Meta-Analysis. *Early Education and Development* 19, 1 (Feb. 2008), 7–26. https://doi.org/10.1080/10409280701838603

[35] Inge Molenaar. 2022. The Concept of Hybrid Human-AI Regulation: Exemplifying How to Support Young Learners' Self-Regulated Learning. *Computers and Education: Artificial Intelligence* 3 (Jan. 2022), 100070. https://doi.org/10.1016/j.caeai.2022.100070

[36] Inge Molenaar. 2022. Towards Hybrid Human-AI Learning Technologies. *European Journal of Education* 57, 4 (2022), 632–645. https://doi.org/10.1111/ejed.12527

[37] Jack Mostow and Wei Chen. 2009. Generating Instruction Automatically for the Reading Strategy of Self-Questioning. *Frontiers in Artificial Intelligence and Applications* 200, 465–472. https://doi.org/10.3233/978-1-60750-028-5-465

[38] Michael Muller and A. Druin. 2012. Participatory design: The third space in human-computer interaction. *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications* (01 2012), 1125–1154.

[39] Mahin Naderifar, Hamideh Goli, Fereshteh Ghaljaie, et al. 2017. Snowball sampling: A purposeful method of sampling in qualitative research. *Strides in development of medical education* 14, 3 (2017), 1–6.

[40] Paul Nation and Robert Waring. 1997. Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy* 14, 1 (1997), 6–19.

[41] Perry Nodelman. 1985. Text as teacher: The beginning of Charlotte's Web. *Children's Literature* 13, 1 (1985), 109–127.

[42] Sylvia Pantaleo. 2017. Critical Thinking and Young Children's Exploration of Picturebook Artwork. *Language and Education* 31, 2 (March 2017), 152–168. https://doi.org/10.1080/09500782.2016.1242599

[43] Michael Pressley. 2002. Metacognition and Self-Regulated Comprehension. In *What Research Has to Say About Reading Instruction* (3 ed.), Alan E. Farstrup and S. Jay Samuels (Eds.). Vol. 1. International Reading Association, Inc., DE, 291–309. https://doi.org/10.1598/0872071774.13

[44] Michael Pressley and Peter Afflerbach. 1995. *Verbal Protocols of Reading: The Nature of Constructively Responsive Reading.* Hillsdale, NJ: Erlbaum.

[45] A Rajagopal and Nirmala Vedamanickam. 2019. New Approach to Human AI Interaction to Address Digital Divide& AI Divide: Creating an Interactive AIplatform to Connect Teachers & Students. In *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT).* IEEE, Coimbatore, India, 1–6. https://doi.org/10.1109/ICECCT.2019.8869174

[46] Deborah Robinson, Nicki Moore, and Catherine Harris. 2019. The Impact of Books on Social Inclusion and Development and Well-being among Children and Young People with Severe and Profound Learning Disabilities: Recognising the Unrecognised Cohort. *British Journal of Learning Disabilities* 47, 2 (June 2019), 91–104. https://doi.org/10.1111/bld.12262

[47] Elizabeth B-N Sanders and Pieter Jan Stappers. 2008. Co-creation and the new landscapes of design. *Co-design* 4, 1 (2008), 5–18.

[48] Kyle-Althea Santos, Ethel Ong, and Ron Resurreccion. [n. d.]. Therapist Vibe: Children's Expressions of Their Emotions through Storytelling with a Chatbot. ([n. d.]).

[49] Margaret Semrud-Clikeman and Kimberly Glass. 2010. The Relation of Humor and Child Development: Social, Adaptive, and Emotional Aspects. *Journal of Child Neurology* 25, 10 (Oct. 2010), 1248–1260. https://doi.org/10.1177/0883073810373144

[50] Sebastian P. Suggate, Wolfgang Lenhard, Elisabeth Neudecker, and Wolfgang Schneider. 2013. Incidental Vocabulary Acquisition from Stories: Second and Fourth Graders Learn More from Listening than Reading. *First Language* 33, 6 (Dec. 2013), 551–571. https://doi.org/10.1177/0142723713503144

[51] Danielle R Thomas, Jionghao Lin, Erin Gatz, Ashish Gurung, Shivang Gupta, Kole Norberg, Stephen E Fancsali, Vincent Aleven, Lee Branstetter, Emma Brunskill, and Kenneth R Koedinger. 2024. Improving Student Learning with Hybrid Human-AI Tutoring: A Three-Study Quasi-Experimental Investigation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference.* ACM, Kyoto Japan, 404–415. https://doi.org/10.1145/3636555.3636896

[52] Daniel Vargas-Diaz, Sulakna Karunaratna, Jisun Kim, Sang Won Lee, and Koeun Choi. 2023. TaleMate: Collaborating with Voice Agents for Parent-Child Joint Reading Experiences. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* ACM, San Francisco CA USA, 1–3. https://doi.org/10.1145/3586182.3616699

[53] Ge Wang, Kaiwen Sun, Ayça Atabey, Kruakae Pothong, Grace C. Lin, Jun Zhao, and Jason Yip. 2023. Child-Centred AI Design: Definition, Operation, and Considerations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* ACM, Hamburg Germany, 1–6. https://doi.org/10.1145/3544549.3573821

[54] Xinghua Wang, Qian Liu, Hui Pang, Seng Chee Tan, Jun Lei, Matthew P. Wallace, and Linlin Li. 2023. What Matters in AI-supported Learning: A Study of Human-AI Interactions in Language Learning Using Cluster Analysis and Epistemic Network Analysis. *Computers & Education* 194 (March 2023), 104703. https://doi.org/10.1016/j.compedu.2022.104703

[55] Zizhen Wang, Jiangyu Pan, Duola Jin, Jingao Zhang, Jiacheng Cao, Chao Zhang, Zejian Li, Preben Hansen, Yijun Zhao, Shouqian Sun, and Xianyue Qiao. 2025. CharacterCritique: Supporting Children's Development of Critical Thinking through Multi-Agent Interaction in Story Reading. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25).* Association for Computing Machinery, New York, NY, USA, Article 131, 21 pages. https://doi.org/10.1145/3706598.3713602

[56] Barbara A. Wasik and Annemarie H. Hindman. 0031/2020. Increasing Preschoolers' Vocabulary Development through a Streamlined Teacher Professional Development Intervention. *Early Childhood Research Quarterly* 50 (0031/2020), 101–113. https://doi.org/10.1016/j.ecresq.2018.11.001

[57] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376781

[58] Xingwen Wu, Peixuan Ye, and Zilin Li. 2021. AI Integration Study into Family Education Guided by Symbolic Interaction Theory: Exemplified by Tmall Genie and Duer in Families of Children Aged 6 to 12. In *2021 2nd International Conference on Education, Knowledge and Information Management (ICEKIM).* IEEE, Xiamen, China, 341–347. https://doi.org/10.1109/ICEKIM52309.2021.00082

[59] Ying Xu, Joseph Aubele, Valery Vigil, Andres S. Bustamante, Young-Suk Kim, and Mark Warschauer. 2022. Dialogue with a Conversational Agent Promotes Children's Story Comprehension via Enhancing Engagement. *Child Development* 93, 2 (March 2022). https://doi.org/10.1111/cdev.13708

[60] Ying Xu, Stacy Branham, Xinwei Deng, Penelope Collins, and Mark Warschauer. 2021. Are Current Voice Interfaces Designed to Support Children's Language Development?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM, Yokohama Japan, 1–12. https://doi.org/10.1145/3411764.3445271

[61] Ying Xu, Stacy Branham, Xinwei Deng, Penelope Collins, and Mark Warschauer. 2021. Are Current Voice Interfaces Designed to Support Children's Language Development?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 633, 12 pages. https://doi.org/10.1145/3411764.3445271

[62] Ying Xu, Kunlei He, Valery Vigil, Santiago Ojeda-Ramirez, Xuechen Liu, Julian Levine, Kelsyann Cervera, and Mark Warschauer. 2023. "Rosita Reads With My Family": Developing A Bilingual Conversational Agent to Support Parent-Child Shared Reading. In *Proceedings of the 22nd Annual ACM Interaction Design and Children Conference.* ACM, Chicago IL USA, 160–172. https://doi.org/10.1145/3585088.3589354

[63] Ying Xu, Valery Vigil, Andres S. Bustamante, and Mark Warschauer. 2022. "Elinor's Talking to Me!":Integrating Conversational AI into Children's Narrative Science Programming. In *CHI Conference on Human Factors in Computing Systems.* ACM, New Orleans LA USA, 1–16. https://doi.org/10.1145/3491102.3502050

[64] Ying Xu, Dakuo Wang, Penelope Collins, Hyelim Lee, and Mark Warschauer. 2021. Same Benefits, Different Communication Patterns: Comparing Children's Reading with a Conversational Agent vs. a Human Partner. *Computers & Education* 161 (Feb. 2021), 104059. https://doi.org/10.1016/j.compedu.2020.104059

[65] Ying Xu and Mark Warschauer. 2019. Young Children's Reading and Learning with Conversational Agents. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–8. https://doi.org/10.1145/3290607.3299035

[66] Ying Xu and Mark Warschauer. 2020. Exploring Young Children's Engagement in Joint Reading with a Conversational Agent. In *Proceedings of the Interaction Design and Children Conference*. ACM, London United Kingdom, 216–228. https://doi.org/10.1145/3392063.3394417

[67] Ying Xu and Mark Warschauer. 2020. What Are You Talking To?: Understanding Children's Perceptions of Conversational Agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376416

[68] Kexin Bella Yang, Vanessa Echeverria, Zijing Lu, Hongyu Mao, Kenneth Holstein, Nikol Rummel, and Vincent Aleven. 2023. Pair-Up: Prototyping Human-AI Co-orchestration of Dynamic Transitions between Individual and Collaborative Learning in the Classroom. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581398

[69] Jing Yao, Xiaoyuan Yi, Xiting Wang, Jindong Wang, and Xing Xie. 2023. From Instructions to Intrinsic Human Values – A Survey of Alignment Goals for Big Models. https://doi.org/10.48550/arXiv.2308.12014 arXiv:2308.12014 [cs]

[70] Hsiu-Ping Yueh, Weijane Lin, S-Chen Wang, and Li-Chen Fu. 2020. Reading with Robot and Human Companions in Library Literacy Activities: A Comparison Study. *British Journal of Educational Technology* 51, 5 (Sept. 2020), 1884–1900. https://doi.org/10.1111/bjet.13016

[71] Andrea A Zevenbergen and Grover J Whitehurst. 2003. Dialogic Reading: A Shared Picture Book Reading Intervention for Preschoolers. *On reading books to children: Parents and teachers* 177 (2003), 200.

[72] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child–AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–15. https://doi.org/10.1145/3491102.3501914

[73] Ke Zhang and Ayse Begum Aslan. 2021. AI Technologies for Education: Recent Research & Future Directions. *Computers and Education: Artificial Intelligence* 2 (2021), 100025. https://doi.org/10.1016/j.caeai.2021.100025

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

[75] Zhao Zhao and Rhonda McEwen. 2022. "Let's Read a Book Together": A Long-term Study on the Usage of Pre-school Children with Their Home Companion Robot. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Sapporo, Japan, 24–32. https://doi.org/10.1109/HRI53351.2022.9889672

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009