How Children Learn with Artificial Intelligence: A Study of Dialogic Story Listening

Astrid Hurtado

Department of Psychology, University of Michigan

April 1, 2024

Author Note

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science with Honors in Biopsychology, Cognition, and Neuroscience from the University of Michigan in 2024

Abstract

As AI becomes more common, its effects on the development of children come into question. The goal of this study is to examine how children learn with artificial intelligence (AI), specifically with an AI voice agent that functions similarly to other popular smart speakers like Amazon's Alexa. Given that dialogic reading has been shown to increase story reading comprehension in children using guided questions, we employed a similar variation of this technique, called dialogic listening, to build story listening comprehension. We were interested in exploring how the benefits of this interactive technique might differ between listening with a human versus with an AI counterpart. To evaluate this, we asked 60 children aged 7-13 to listen to a third-grade reading level story, *Henry Huggins*, and answer questions asked by either a human conversational partner or AI voice agent. The children's responses were evaluated by analyzing five subcomponents of verbal engagement: language productivity, lexical diversity, topical relevance, the accuracy of the response, and intelligibility of the child's utterances. While our analysis failed to demonstrate similarities in overall story listening comprehension across human and AI conditions, children exhibited comparable levels of verbal engagement and even responded to the AI voice agent with fewer errors. In sum, our findings suggest that while AI is not yet as effective in scaffolding story comprehension as a human, it is able to engage children despite limitations to its current design.

Keywords: artificial intelligence, story comprehension, early literacy skills, dialogic listening, verbal engagement

Acknowledgments

This research has been generously supported by the Tanner Memorial Award, which was provided by the family of Wilson P. “Spike” Tanner. Thanks to the funding from this award, I was able to meet my initial goal of incorporating 60 participants and ensure quality data by recruiting additional help to streamline data collection.

I’d like to thank my Principal Investigator and mentor, Dr. Ioulia Kovelman, who supported me through the entirety of this process all while simultaneously overseeing several other on-going projects in the lab and 20 something other research assistants. Thank you to my co-mentor, Dr. Ying Xu, whose research helped shape the framework of our study. I also want to extend a special thank you to my graduate student mentor, Chi-Lin Yu, who generously lent me his invaluable expertise and guided me through the complex data collection and analysis process, always encouraging me to think outside the box and constantly finding ways to improve our research. I’d also like to thank my lab manager, Isabel Hernandez, who was the first person who encouraged me to write a thesis. Thank you for always giving me kind words of encouragement when I needed them the most. Finally, I’d like to thank my family and especially my mom, whose work has inspired me to follow her footsteps into the world of research.

How Children Learn with Artificial Intelligence: A Study of Dialogic Story Listening

Despite having only been in existence for the past 50 years, artificial intelligence (AI) has become increasingly accessible to the public, often in the form of smart devices (Buchanan, 2005). According to a recent study conducted by Kinsella et al. (2020), more than 150 million households in the U.S. owned smart speakers in early 2020. With over half of the American population owning smart speakers, and the number projected to increase in the following years, smart devices have become an integral part of daily life for young children growing up with them, both in and out of the home.

The rising availability of AI correlates with the growing trend of implementing technology as a learning tool in early childhood education (Gampe et al., 2023). As AI becomes more readily available to the public, its influence on the development of young children, at home, and school, has become a pressing question in the field of psychology. Although smart devices have become increasingly common, there is a substantial lack of research regarding the role of AI in the context of language development. Thus, this study aims to explore how young children interact and learn with AI through dialogic listening and will consider their verbal engagement, level of story comprehension, and the types of questions asked during their session.

How Children View Artificial Intelligence

Prior research demonstrates that children tend to anthropomorphize and readily engage with AI voice agents, indicating the potential of AI as a learning tool that models established human-scaffolded techniques. In their study, Lee & Jeon. (2022) noted that most child participants did not view the AI voice agent as a human. Instead, they considered it as a mixture of both. Children's willingness to engage with AI on a more personal level, compared to adults, is promising because children exhibit higher levels of reading engagement and comprehension when they bond or have a positive relationship with their conversational agent (Gampe et al., 2023). This finding points towards the possibility of AI as an effective interactive conversational partner for language development (Lee & Jeon., 2022).

Storybook Reading and Emergent Literacy

The use of AI as a learning tool for young children has been previously investigated by researchers in the context of children's reading comprehension following both storybook and dialogic reading techniques (Lee & Jeon., 2022; Xu et al., 2021). Storybook reading is an activity that involves the conversational agent, usually a parent, sitting with and reading picture books to the young child (Xu et al., 2021). For children who are not yet able to read or write, generally under six years of age, storybook reading is an activity that develops important early or emergent literacy skills (Strouse et al., 2018).

While literacy refers to the ability to read and write, emergent literacy is a developmental precursor to reading and writing that consists of foundational skills that include phonological awareness and letter knowledge, which are developed through storybook reading (Whitehurst & Longian, 1998). As a precursor to literacy, emergent literacy can be used to predict or gauge future literacy in preschool-aged children who are not yet reading and writing (Whitehurst & Longian, 1998). This makes activities that foster emergent literacy, such as storybook and dialogic reading, significant in the context of how young children develop language and learn to read.

The Effect of Dialogic Reading on Reading Comprehension

Dialogic reading is a more recently developed interactive reading style that builds upon the foundations of storybook reading (Kleek & Whitehurst, 2009, p.171). This reading technique employs scaffolded adult-child interactions to stimulate children's thinking and provide feedback, increasing children's overall engagement (Kleek & Whitehurst, 2009, p.171; Xu et al., 2021). While both storybook and dialogic reading support the development of early literacy skills, dialogic reading, which intersperses questions and conversation throughout the story, is more beneficial for building story comprehension compared to storybook reading (Xu et al., 2021). According to Strause et al. (2018), having engaging conversations is the most important thing that adults can do to help children learn, especially when they are learning how to read.

Xu et al. (2021) further demonstrate the positive impact of dialogic reading by considering how story comprehension varies depending on the conversational partner children interact with, either an AI voice agent or a human counterpart. The researchers tested four conditions, alternating dialogic reading and non-dialogic (storybook) reading strategies, with either a human or an AI voice agent that followed the same script throughout. Xu et al. (2021) found that only dialogic reading, not the type of conversational partner, had a significant effect on story comprehension. Children exhibited higher levels of engagement and narrative-relevant vocalizations when dialogic reading was used by the AI voice agent. There was a 0.51 standard deviation increase in the children's story comprehension score when dialogic reading was employed, regardless of whether children interacted with a human partner or an AI conversational agent (Xu et al., 2021). Given that reading engagement predicts the growth of story comprehension, this finding suggests that AI could *potentially* be as effective as a human counterpart when it comes to scaffolding story comprehension in young children through dialogic reading. (Barber et al., 2020; Xu et al., 2021).

The Effectiveness of Artificial Intelligence as a Learning Tool

The efficacy of AI as an instrument for learning has been debated by researchers in the field, in part due to a lack of specific parameters outlining what constitutes as an effective instructional resource. In their study, Xu et al. (2021) expanded on the relevance of the contents of the conversational partner's script during dialogic reading by considering both the accuracy and length of the children's responses to different prompts in addition to the difficulty of the questions asked. Children read to by an adult had greater lexical diversity, relevance, and higher productivity in their responses compared to children read to by an AI voice agent (Xu et al., 2021). Although the accuracy of the children's responses did not significantly differ between conversations with an AI voice agent compared to with a human, the *quality* of the response did, suggesting that AI might be used as a supplementary tool for preschool children, but that AI is not yet capable of replacing vital parent-child interactions (Xu et al., 2021).

Additional research has demonstrated that children employ different repair strategies when communication breaks down with an AI voice agent compared to with a human (Gampe et al., 2023). Gampe et al. (2023) found that children took longer to respond to and were less likely to attempt to fix misunderstandings with an AI voice agent, highlighting that AI might benefit from incorporating more engaging acoustic features, such as tone and inflection, to better engage children (Gampe et al., 2023). Additionally, by incorporating dialogue into the script to mimic human conversation, AI might be a more effective tool for language development in young children (Xu et al., 2021).

Limitations of AI speech-based agents

Although research has demonstrated that AI has the potential to be an effective tool for literary acquisition, current AI voice agents are limited by their monotone voices and basic scripts (Gampe et al., 2023). Their limited conversational ability prevents them from conveying the necessary emotions, like excitement and surprise, needed to maintain the child's interest in the topic. More engaged children make stronger attempts to resolve misunderstandings when they have a stronger bond with their conversational agent (Gampe et al., 2023). Given that engagement correlates with greater story comprehension, the limited script could potentially hinder overall comprehension of the story or lesson.

During dialogic reading, adults often engage children in the story by drawing their attention back to content-related talk and are better able to tailor their questions to the child, which are features that AI does not yet have (Strouse et al., 2018). While some studies have employed scripts with guided questions designed to promote a similar scaffolding effect, the pre-determined responses are limited and eventually move on to the next question regardless of the child's answer (Xu et al., 2021). Additionally, while adult humans use verbal gestures and social cues to prompt children during their conversations, AI voice agents, by design, are limited to verbal cues, hindering their ability to adequately convey their expectations in the same manner that a human might, through body language or facial gestures (Okon, 2011).

Types of Questions Asked during Dialogic Reading Matter

Although it is widely accepted that dialogic reading is the most effective literary technique for scaffolding reading comprehension and sustaining children's engagement, no consensus has been reached regarding the influence of the type of question asked during the session. While one study has established that asking more complex, cognitively demanding, questions results in greater reading engagement, no significant difference in intelligibility or accuracy was found when comparing children's performance across high or low cognitive demand questions (Xu et al., 2021). In contrast, open-ended questions, which require the participant to reflect and assess more complex factors, have been shown to require greater recall and narrative elaboration, resulting in a lower rate of accuracy compared to multiple-choice type questions that have predetermined answers participants select (Ozuru et al., 2013). Given the broad range in the type of question that can be asked during dialogic reading, it is difficult to establish a clear pattern specifying the benefits of one type of question over the other.

Additionally, there is limited research regarding the effects of questions that prompt children's recursive thinking ability, inferential questions that test their ability to "mind-read." Theory of Mind is a concept that recently emerged in the 1970s and describes the ability to consider the thoughts or mental states of others (Beaudoin et al., 2020). It first appears in its most basic form around four years of age, developing throughout childhood and adolescence (Valle et al., 2015). The development of theory of mind marks a cognitive milestone essential for social interactions and advanced reasoning, making it important to consider in the context of early language development.

The present literature indicates that some types of questions, such as open-ended and multiple-choice questions, are answered by activating different processes of comprehension (Ozuru et al., 2013). While the underlying neural mechanisms are yet to be fully understood, this variation suggests that different questions could potentially be used to scaffold different skills. Perhaps asking different types of questions during dialogic reading might be more beneficial than only asking one type of question, making the type of question asked a relevant factor in the context of story comprehension. Because little is known about the effect of inferential questions, it is interesting to consider how they might interact with the way that children understand a story in the presence of an AI voice agent. This study aims to build upon prior research by exploring how questions that cue theory of mind influence children's verbal engagement and story comprehension.

Present Study

Overall, this study primarily explores the nuances of literary acquisition in children within the context of dialogic listening facilitated by AI, asking, “How do children learn with AI?” Within this broader context, I ask two more specific questions, “Do children learn as effectively with AI compared to with humans?” and “How do the narrative responses of children vary in the presence of AI?” To answer these questions, I analyzed children’s verbal responses during a story-listening task to evaluate their story comprehension and verbal engagement.

I hypothesized that children would demonstrate a comparable level of overall story comprehension with either the human or AI-voice agent, but that they would produce less-complex responses when conversing with the AI voice agent. The AI voice agent’s inability to use nonverbal gestures to convey its expectations might hinder its ability to prompt the child as well as a human counterpart, resulting in less engagement and shorter responses overall. I also predicted that participants would produce longer, more complex narrative answers, with a higher frequency of filler pause words, in response to mental questions as well, which require access to one’s mental state, as compared to non-mental questions that simply require an understanding of the physical world of the story.

Methods

Participants

This study includes data from 60 English-speaking children ranging from 7 to 13 years of age, all of whom are from the Midwest. All children fell into the middle-class demographic. A background questionnaire was filled out by the parents of the participants to gather additional demographic and language data prior to the experiment (see Figure 2). This questionnaire included information such as the parent’s level of education, any speech or language impairments, and the language spoken at home, which in this case was primarily English. Only native English speakers were included in this study. These metrics were relevant to the study because they allowed us to consider any potential cognitive or physical impairments when administering the behavioral assessments toward the end of the testing session.

Overview of Experimental Design

The experimental setup consists of a 2 (human conversational agent vs. AI voice agent) x 2 (mental questions vs. non-mental questions) factorial design, as illustrated by Figure 1. 30 children were paired with the AI voice-agent, while the other 30 children were paired with the human conversational partner. The order of assessments underwent by the child participants is visualized in Figure 2. Overall, each testing session ranged between three and four hours. Each testing session consisted of a story-listening activity, followed by a post-reading comprehension task, and ended with behavioral testing comprised of several literacy assessments to ensure participants fell within a standard range of reading and reasoning ability across both conditions.

Figure 1

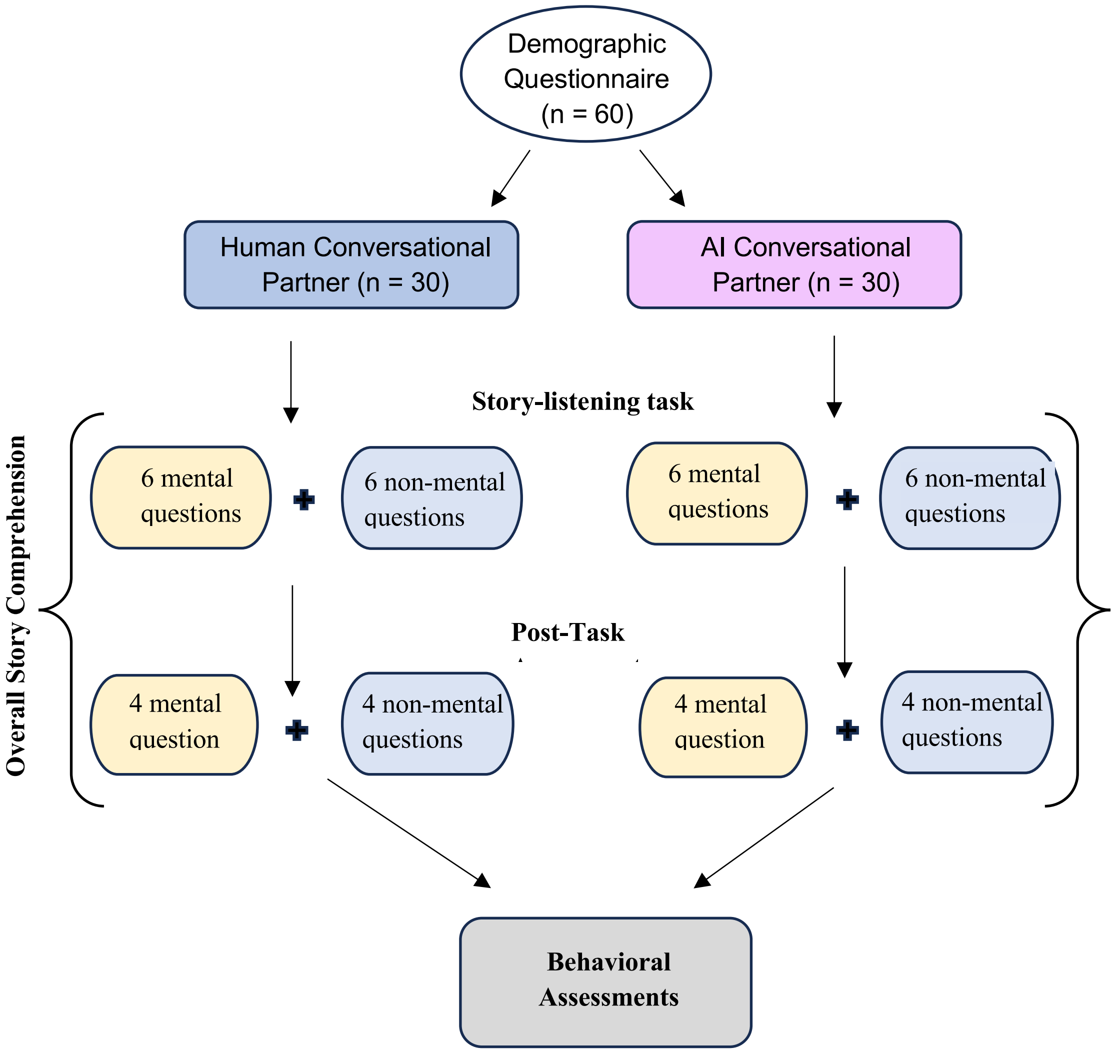
2x2 Factorial Design Set-Up

		The Effect of the Type of Question	
		Mental Questions	Non-mental Questions
The Effect of the Conversational Partner	Human	asks questions that require theory of mind.	Asks questions about the physical world of the story.
	AI	Asks questions that require theory of mind.	Asks questions about the physical world of the story.

Note. Participants in the human and AI conditions (*N* = 60) were asked mental and non-mental story comprehension questions during the in-task and post-task assessments. In total, participants were asked 20 questions, 10 of which were mental and 10 which were non-mental, in a random order.

Figure 2

Flowchart of Testing Session Protocol



Note. Participants (N = 60) were equally sorted into both the human and AI conditions. After parents filled out the demographic questionnaire, all participants were asked the same 20 questions in a random order during both the in-task and post-task portions of the session. All behavioral assessments were conducted by a human examiner at the end of the testing session.

Story-listening Task

Children engaged in a 20-minute-long story-listening task, in which either an AI voice agent or a human conversational partner asked them questions about the events of the story to scaffold their understanding. In order to adequately evaluate the potential of AI as an educational resource, we chose dialogic listening as the framework for our study given the established efficacy of dialogic reading as a tool for literary acquisition. The main difference between these techniques is that dialogic listening involves an audio-book, allowing us to evaluate story listening comprehension, while dialogic reading uses a physical story book and builds story reading comprehension. Both activities implement questions throughout the story to engage the child. To maintain consistency, a human experimenter recorded the script used by the AI voice agent, minimizing the potential impact of the tone of voice, as stiff, monotone robotic voices can negatively impact story comprehension (Gampe et al., 2023).

During the audio recording of *Henry Huggins*, a chapter book written at the 3rd grade reading level, 12 questions were asked by the assigned conversational partner throughout the story, in no particular order, as part of the in-task comprehension assessment. The questions asked were classified as either mental or non-mental (see Figure 1). Both mental and non-mental questions share a fact-type structure, meaning that their answers can be directly derived from the content of the story. The main difference between the two is that mental questions target theory of mind. Of the 12 questions asked, six questions were classified as mental questions, which require the participant to consider the mental states of the characters in the story such as, the protagonist, Henry Huggins, his mom, and even the other passengers on the bus. An example of a mental question asked is, “Why did Henry think that was a good time to hang up?” This question requires the participant to think about Henry’s motivation for ending the call, requiring theory of mind to correctly answer the question. The other six questions asked were non-mental questions, which are questions about the physical world of the story. An example of this type of question is, “What was the size of the dog?” None of the possible answers to this question, which are variations of small, medium, or large, require theory of mind. It is important to note that neither mental nor non-mental questions correlate to the level of cognitive demand, but rather to theory of mind, or the lack thereof.

The post-task comprehension assessment was administered immediately following the conclusion of the story-listening task. Unlike the story-listening task, verbal engagement was not measured, and no audio-visual recordings were collected during the post-task assessment; this task was solely conducted to evaluate story comprehension through eight additional questions. Similarly to the in-task assessment, the post-task consisted of both mental and non-mental questions, but in this case, there were only four mental questions and four non-mental questions. All post-task assessments were administered by a human experimenter, regardless of whether the child previously interacted with a human or AI voice agent. The inclusion of the post-task ensured that the present study accurately captured the participants’ overall comprehension of the story.

Behavioral Testing

Participants were asked to complete English language literacy assessments as part of their behavioral testing following the story listening task. These were conducted in the following order: Early Lexical Morphology Measure (ELMM; Marks et al., 2022), Comprehensive Test of Phonological Processing (CTOPP-2; Wagner et al., 2013), passage comprehension (WJ-PC; Schrank et al., 2014), Letter Word Identification (WJ-LWID; Schrank et al., 2014), Test of Word

Reading Efficiency (TOWRE; Wagner et al., 2013), and the Kaufmann Brief Intelligence Test (KBIT-2; Kaufman & Kaufman, 2004, as cited in Allen, 2013).

All these tests measure components that are important precursors to early literacy skills. ELMM measures the participants' morphological awareness, which reflects an understanding of how different parts of a word affect its meaning. CTOPP measures phonological awareness, which is the ability to recognize different sounds in parts of words. These are both early literacy skills that impact word reading and comprehension (Whitehurst & Longian, 1998). Passage comprehension measures how well the child comprehends what they read in a series of increasingly longer and more complex passages, directly testing their ability to use context clues to fill in the missing word of the passage. LWID measures how well the child recognizes and pronounces words, assessing their oral reading skills. KBIT measures both verbal and nonverbal intelligence, two components of IQ that relate to reasoning ability and overall cognitive skills. KBIT is split up into a picture identification test that is followed by a riddle test that subsequently increases in difficulty as the child progresses.

The behavioral assessments were conducted to ensure that children assigned to both groups, either with the human conversational partner or AI voice agent, had a similar reading ability to reduce the potential impact of language and cognitive skills as a covariate. This ensured that differences observed in story comprehension and verbal engagement were due to how successfully either the human or AI voice agent scaffolded the child's understanding of the story, not other language-related confounding factors.

Transcript Conventions

The 20-minute audiovisual recordings were edited down to the question-answer segments of the story-listening task and auto-transcribed into a written transcript using Whisper AI (2023).

This raw transcript was then segmented, behaviorally coded, and analyzed using SALT, Systematic Analysis of Language Transcripts, (MAC Version 20; Miller, J. & Iglesias, A., 2020) software. In accordance with SALT conventions, the phrases of the conversational partner, the human or AI voice agent, and the child participant were segmented into individual communication units (C-units). Each C-unit refers to an independent clause and its modifiers, which include linked sentence fragments and dependent clauses (Version 20; Miller, J. & Iglesias, A., 2020). Adhering to the segmentation guidelines was essential to ensure that key components such as the number of contractions and past-tense words spoken, were accurate. Inaccurate segmentation of these C-segments would skew the calculated length and complexity of the sample.

Following segmentation, transcripts were behaviorally coded using the conventions outlined by SALT. The extensive process included coding for bound morphemes, mazes, omissions, overlapping speech, contractions, past-tense words, errors, etc. Once all specific parts of speech were coded, the main 12 question transcript was subdivided into two transcripts: a six-question transcript comprised of mental questions and another six-question transcript composed of non-mental questions. This allowed us to consider the effect of the type of question asked in addition to that of the AI voice agent. In sum, all participants were assigned two transcripts, making the total number of transcripts analyzed 120. It is important to note that SALT was only used to analyze the participants' verbal engagement recorded during the story-listening task, not story comprehension or the participants' responses during the post-task assessment.

Evaluating Story Comprehension

Table 1 shows both the average in-task and post-task story comprehension. Story comprehension was measured by the accuracy of the children's responses to the questions asked by the conversational partner, either human or AI. Story comprehension was measured by the in-task and post-task assessments, which were administered during and immediately after the story-listening task, respectively. Both the in-task and post-task assessments were binomially scored, meaning they were marked as either correct (1) or incorrect (0). In order to observe the effect of the type of question asked, story comprehension was separated across mental and non-mental questions. Regardless of the conversational partner assigned, each participant received two story comprehension scores, with a maximum score of 6 for the in-task and 4 for the post-task. This accounted for the respective six mental and non-mental questions asked during the in-task assessment as part of the story-listening task and the four mental and non-mental questions asked during the post-task assessment.

Metrics for Verbal Engagement

The children's verbal engagement was measured by the quality of their responses. Given the broad, almost intangible, definition of what a "quality response" is, this variable was operationalized by analyzing five subcomponents thought to be critical reflections of the participants' language proficiency and level of verbal engagement. These included: language productivity, lexical diversity, topical relevance, accuracy of the response, and intelligibility of the child's utterances (Xu et al., 2021).

Language Productivity

Language productivity, which generally refers to the number and length of phrases produced by the speaker, was measured by evaluating the Mean length of utterance (MLU), the percentage of abandoned utterances, the number of total words spoken (NTW), the number of words were spoken per minute, and pause time percentage. A morpheme is the smallest part of a word with meaning. Because MLU measures the average number of morphemes in an utterance, a higher MLU value indicates longer more complex utterances and greater language productivity. The percentage of abandoned utterances marks how frequently the child trailed off in their response and either changed topic or concluded their thought. The percentage of abandoned utterances refers to the time a child spent paused or thinking as a percentage of the total time spent conversing with their conversational partner. The number of words spoken per minute indicates the speed at which the child spoke; a greater number of words per minute equates to a faster rate of speech.

Lexical Diversity

Lexical diversity, which refers to the number of unique words spoken, was measured using Type Token Ratio (TTR). TTR is the ratio between the number of different words spoken (NDW) and NTW, ranging from 0 to 1. This means that the more unique words a speaker produces, the greater the TTR value, the closer it is to 1, and the more lexically diverse their response will be. It was important to consider this metric in conjunction with language productivity, as a longer response, indicated by a higher NTW, is not necessarily more lexically diverse and could just contain more of the same words.

Topical Relevance

Topical relevance refers to how closely the participant's responses related back to the story-related content of the original question asked. This was evaluated by comparing both the percentage of maze words and the number of filler pause words produced overall. A maze is defined as a non-essential utterance that includes repetitions and false starts (Version 20; Miller, J. & Iglesias, A., 2020). While maze words are simply all the words that form a maze, filler pause words are specific non-essential words within a maze that include variations of "um," "mmm," and "like" (Version 20; Miller, J. & Iglesias, A., 2020). For this reason, filler pause words were considered as a special subset of the mazes produced by the speaker; filler pause words within a maze were assigned the code [FP] for additional analysis. The greater number of filler pause words and mazes contained within a sample, the less topically relevant the speaker's response will be, as the response contains a greater proportion of words and phrases that are irrelevant to the intended meaning.

Accuracy

Accuracy was measured by the percentage of utterances containing errors. Utterance-level errors marked large-scale syntactical errors that made the meaning of the child's utterance impossible to discern. Utterances were assigned error code [EU] when there were more than three errors concerning the grammatical structure of the entire utterance or phrase. For example, "And they went to stopped" is a phrase that would be marked as an utterance-level error, as its faulty syntax makes it difficult to understand the child's original intent and assign individual word-level errors. A higher the frequency of utterance-level errors results in a lower accuracy overall.

Intelligibility

Finally, intelligibility of the children's responses was measured by the percentage of utterances that were understood by the experimenters. Unlike utterance-level errors, which relate to improper syntax, the specific words spoken within unintelligible utterances were not easily understood or agreed upon by the experimenters. When an utterance was indecipherable by both coders, the phrase was considered "unintelligible" and assigned the error code [XXX]. By default, any response without the error code [XXX] was considered "intelligible." A higher percentage of intelligibility equates to a more easily understood response overall, consisting of less unintelligible utterances marked [XXX].

Statistical Analysis

A mixed two-way repeated measures analysis of variance (ANOVA) (story type X group type) analysis was run using the Real Statistics Resource Pack for Microsoft Excel (Version 8.9.1; Zaiontz, 2023) across all verbal components and different metrics of story comprehension. This allowed us to compare the children's performance in response to both mental and non-mental questions asked across the two main experimental groups, the human and AI-voice agent conditions. The ANOVA test was run for both measures of story comprehension and all 10 individual metrics listed in Tables 1 and 2, respectively.

Two coders were used to ensure the reliability of the results; both of whom are native English speakers. The primary coder edited, transcribed, and behaviorally coded all 60 original audiovisual samples, which yielded 120 transcripts. The second coder checked one third of all transcripts by listening back to the audio for each question and verifying that proper notation was used. All checked transcripts were re-run through SALT and incorporated into the final data analysis.

Results

Summary Data

Table 1 summarizes the story comprehension of participants on the in-task and post-task assessments. Table 2 outlines the 10 SALT metrics used to quantify verbal engagement and the participants' respective performance. The contents of both tables depict the calculated mean and standard deviation values across the human and AI conditions for the children's responses to both mental and non-mental questions asked during the testing session.

Figure 3 compares the overall language skills between the participants assigned to the human and AI conditions. As expected, in terms of performance on the six major behavioral assessments, there was no significant difference between participants who interacted with a human conversational partner and those who interacted with the AI voice agent (*ps* > .05). Since all 60 child participants exhibited a similar baseline in terms of general language ability, any additional differences observed in performance during the story-listening task and post-task assessment are likely due to either the conversational partner assigned, or the type of question asked, rather than their general language ability.

Table 1

Story Comprehension Measured Across Human and AI Conditions

Type of Assessment	Mental Questions		Non-mental Questions		*p-value* (between)
	Human M (SD)	AI M (SD)	Human M (SD)	AI M (SD)	
% in-task story comprehension	87.8 (13.8)	66.7 (19.3)	92.8 (12.9)	80.6 (18.8)	< .001
% post-task story comprehension	94.2 (12.6)	96.8 (8.5)	98.3 (6.34)	92.7 (13.2)	.44

*p < .05.

Note. Table 1 summarizes the mean and standard deviations of the percentage of questions participants (*N* = 60) got correct on both comprehension assessments, which were scored out of 6 and 4 points, respectively. All listed *p-values* compare performance between the human and AI assigned groups.

Table 2

Verbal Engagement Measured Across Human and AI Conditions.

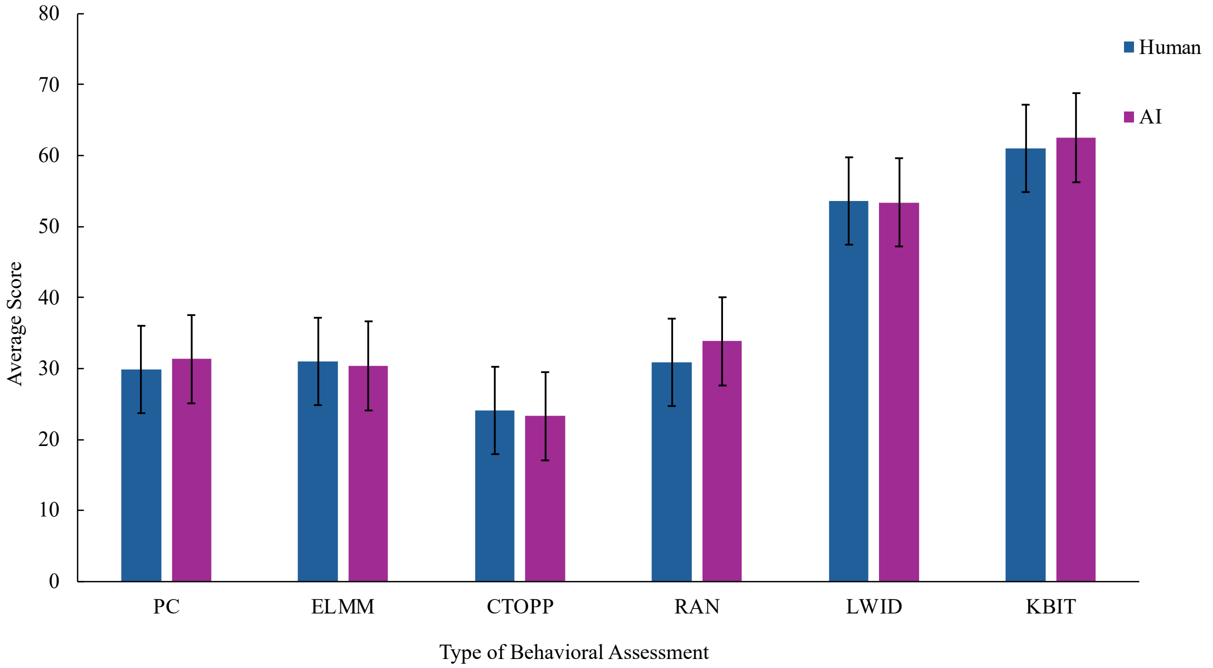
Verbal Facility	SALT Metrics	Mental Questions		Non-mental Questions		*p-value* (between)
		Human M (SD)	AI M (SD)	Human M (SD)	AI M (SD)	
Language Productivity	MLU	8.7 (2.2)	8.2 (2.9)	5.6 (1.9)	5.1 (1.9)	.36
	% abandoned utterances	2.9 (5.4)	0.5 (2.6)	1.2 (8.0)	1.0 (3.9)	.10
	NTW	68.5 (29.2)	53.1 (22.8)	41.5 (17.8)	31.8 (14.4)	.01
	Words per minute	42.7 (13.1)	27.9 (11.0)	31.3 (12.8)	21.2 (10.0)	< .001
	Pause time %	2.4 (3.6)	4.5 (6.5)	3.6 (5.5)	4.7 (9.5)	.36
Lexical Diversity	TTR	0.7 (0.1)	0.7 (0.1)	0.7 (0.1)	0.8 (0.1)	.15
Topical Relevance	% of maze words	8.6 (6.6)	4.9 (4.7)	9.2 (5.9)	5.9 (7.0)	.01
	Filler pause words	2.6 (2.5)	1.2 (2.0)	2.0 (2.0)	1.0 (1.8)	.01
Accuracy	% utterances with errors	16.2 (13.6)	7.2 (10.7)	6.9 (8.0)	1.4 (4.3)	< .001
Intelligibility	% intelligibility	98.0 (5.2)	99.5 (3.0)	98.4 (4.3)	100 (0.0)	.02

**p* < .05.

Note. Table 2 summarizes the mean and standard deviations of the participants' (*N* = 60) performance on 10 SALT-derived metrics based on the type of question asked by the human or AI voice agent. All listed *p-values* compare performance between the human and AI assigned groups.

Figure 3

Average Language Ability of Participants Assigned to Human and AI Conditions



Note. Figure 3 shows the average raw scores for all six behavioral assessments administered by a human, which evaluated different aspects of the participants' (*N* = 60) language and cognitive skills. Error bars represent standard error.

Story Comprehension

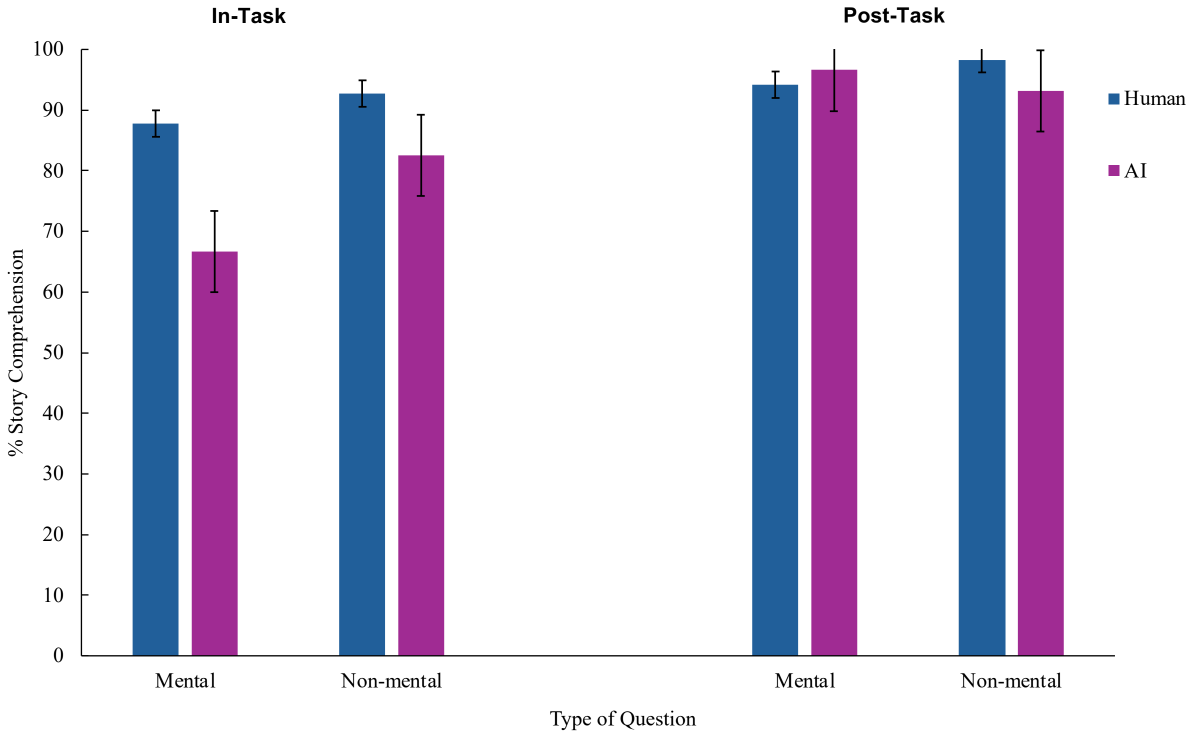
Our main question was whether children would be able to accurately comprehend a story after interacting with the AI voice agent. Story comprehension was primarily evaluated through the percentage of questions that participants got correct on both the in-task and post-task assessments as shown by Figure 4. Table 1 summarizes the average scores of participants across the four sub-categories in which story comprehension was compared during both comprehension assessments. The four categories, which were derived from the assigned conversational partner and type of question asked are as follows: human/mental, AI/mental, human/non-mental, and AI/non-mental.

The mixed two-way repeated measures ANOVA test revealed an interaction between the effect of the conversational partner assigned and type of question asked, specifically in the post-task story-comprehension assessment (*p* = .02). Follow-up data analysis revealed that children responded to non-mental questions asked by the human conversational partner with greater accuracy overall compared to children who interacted with the AI voice agent (see Table 1; *p* = .02). Conversely, children exhibited greater accuracy for mental questions in response to the AI voice agent compared to those with the human conversational partner. No significant trend was observed in participants' responses to mental questions; children in both the human and AI conditions exhibited comparable story comprehension, with marginally greater accuracy observed in response to the human conversational partner (*p* = .19).

Unlike the post-task assessment, no interaction was observed in terms of the children's performance on the in-task assessment. However, a main effect for the assigned conversational partner, human or AI, was observed. Overall, children paired with the AI voice agent exhibited a lower degree of story comprehension compared to children who were paired with the human conversational partner. Additionally, a main effect was observed in terms of the type of question asked within participants assigned to the human and AI conditions. Overall, children responded to non-mental questions with greater accuracy. While no interaction was observed (*p* = 0.09), participants generally answered mental questions asked by the AI voice agent with a lower level of story comprehension compared to participants who were paired with the human conversational partner. A similar marginal trend was observed when considering non-mental questions; participants answered non-mental questions asked by the AI voice agent with less accuracy than participants in the human condition.

Figure 4

Comparison of Story Comprehension Across Type of Question and Conversational Partner.



Note. Figure 4 compares the percentage of question participants' (*N* = 60) answered correctly in response to either mental or non-mental questions asked by their assigned conversational partner. Error bars represent standard error.

Verbal Engagement

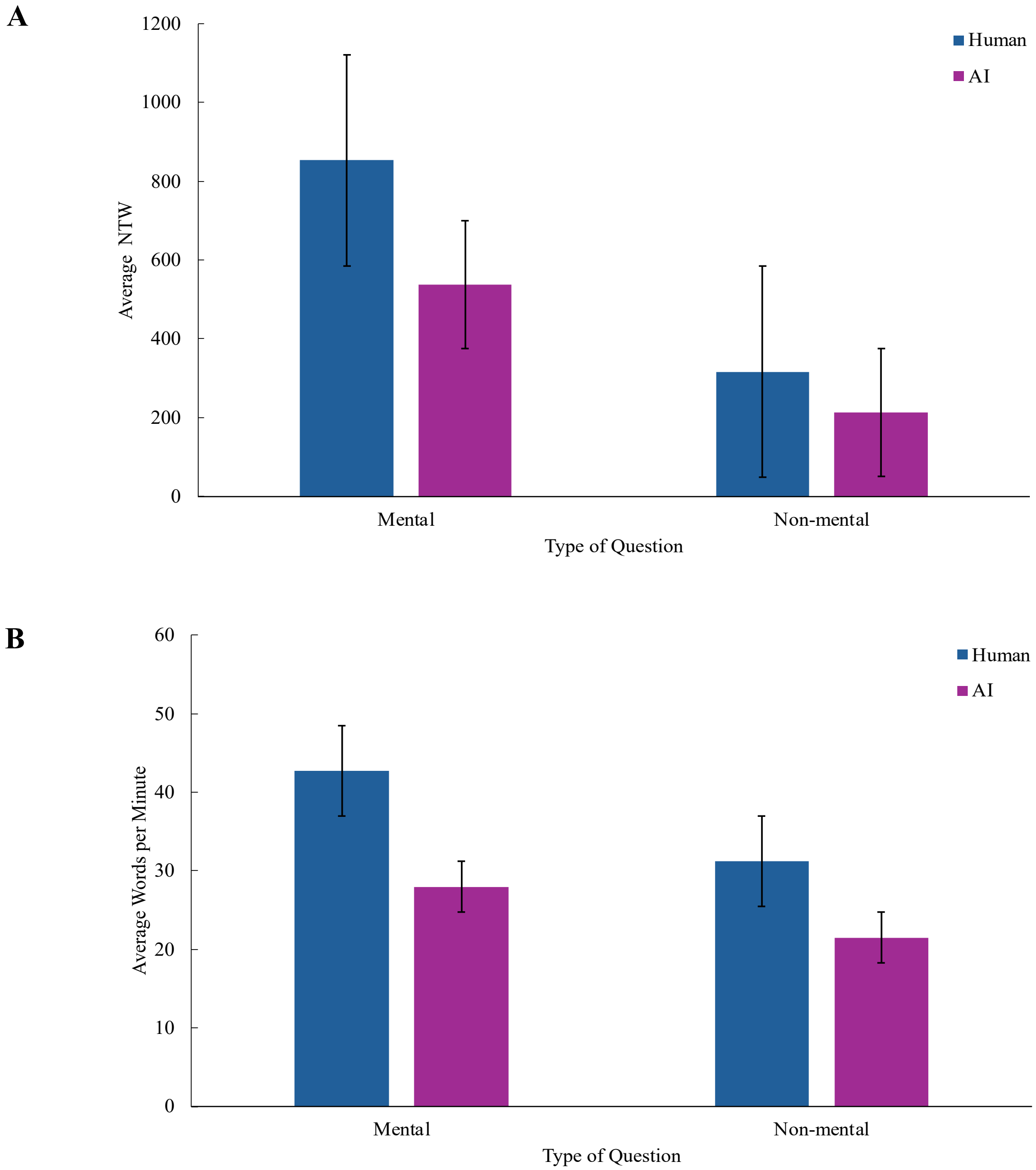
Language Productivity

Four of the five metrics used to evaluate language productivity – MLU, the percentage of abandoned utterances, NTW, and the number of words spoken per minute – reflected the general trend of children using more expressive language in response to the human conversational partner (see Appendix A). Of these four metrics, only NTW and words per minute, exhibited an additional main effect for the type of question asked, as shown in Figure 5. On average, children who interacted with the human conversational partner spoke with a greater NTW. Additionally, compared to non-mental questions, children responded to mental questions with a greater NTW overall. Children were also found to speak faster, by producing more words per minute, in response to the human conversational partner. Similarly to the trend observed in the NTW produced, children responded faster to mental questions with a greater number of words per minute as well.

No main effect for the type of question asked was observed for MLU, the percent of abandoned utterances, and the percentage of time spent paused. While no main effect was established between children in the human or AI conditions either, children who interacted with the human conversational partner exhibited a marginally higher MLU compared to children who were paired with the AI voice agent. Likewise, children paired with the human conversational partner had a greater proportion of abandoned utterances compared to children in the AI condition. Finally, when evaluating the percentage of time that the participants were paused, when they were not actively responding to the examiner, no significant difference was found when comparing performance between the human and AI conditions. However, a marginally significant trend was observed, indicating that participants took longer to answer questions asked by the AI voice agent.

Figure 5

Comparison of Language Productivity



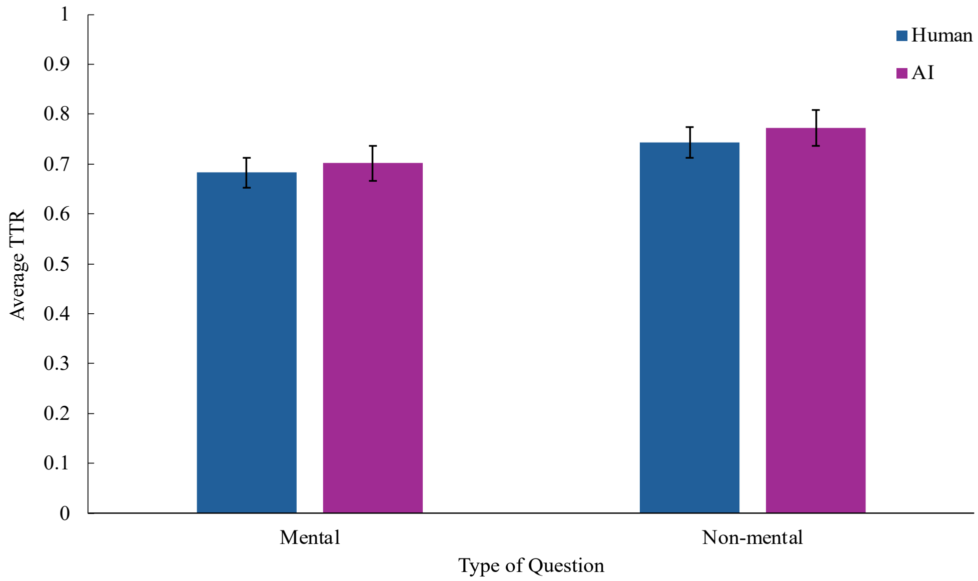
Note. (A) Average NTW & (B) Average Words per Minute. Figures 5A and 5B show the average number of total words and speed at which participants (*N* = 60) spoke based on the type of question asked by their assigned conversational partner, human or AI. Error bars represent standard error.

Lexical Diversity

A main effect of the type of question asked within the human and AI conditions was observed, as shown by Figure 6. In general, participants responded to mental questions with a lower degree of lexical diversity (i.e., a lower average TTR) compared to when they were asked non-mental questions. This means that children's responses to mental questions contained a lower proportion of unique words compared to when they responded to non-mental questions. No other effects were found.

Figure 6

Comparison of Lexical Diversity



Note. Figure 6 shows the participants' (*N* = 60) average TTR scores, which are categorized by the human or AI conversational partner and type of question asked. TTR scores ranged from 0 to 1, with 1 representing a maximally diverse sample. Error bars represent standard error.

Topical Relevance

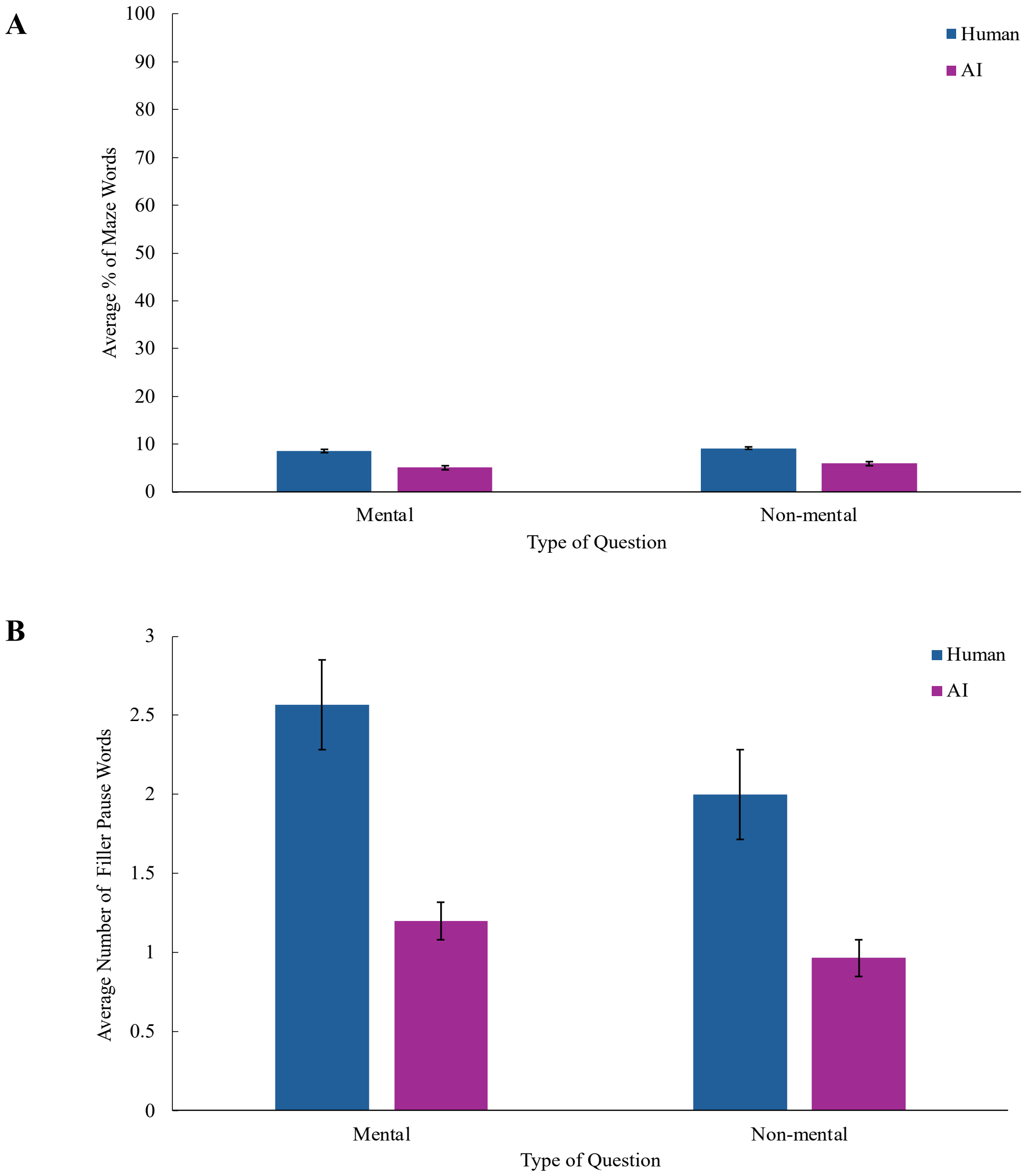
Topical relevance, which considers how closely related the participants' responses aligned with the accepted answer, was analyzed by assessing both the percentage of maze words in each sample and the average number of filler pause words. As shown by Figure 7, participants generally produced topically relevant responses, with maze words only contributing to around 10% or less of children's responses across both groups. On average, child participants produced more mazes when interacting with the human conversational partner but did not exhibit a significant effect in terms of the type of question asked. Similarly, the responses of participants who interacted with the human conversational partner contained, on average, more filler pause words compared to those with the AI voice agent. No significant difference was observed in terms of the effect of the question asked. The greater number of maze words and filler pause words produced in response to the human conversational partner illustrates that children overall responded to the AI voice agent, instead of the human, with more topically relevant responses.

Accuracy

While no interaction was observed, the ANOVA test yielded significant effects for both the assigned conversational partner and type of question asked on the percentage of utterances containing errors. Overall, children produced a greater percentage of utterance-level errors in response to the human conversational partner compared to children who interacted with the AI voice agent, as displayed by Figure 8. In terms of the effect of the type of question asked, child participants displayed a significantly higher rate of errors in response to mental questions ($p <$ .001). Additionally, a non-significant correlation was found indicating that children produced the lowest number of errors when answering non-mental questions asked by the AI voice agent.

Figure 7

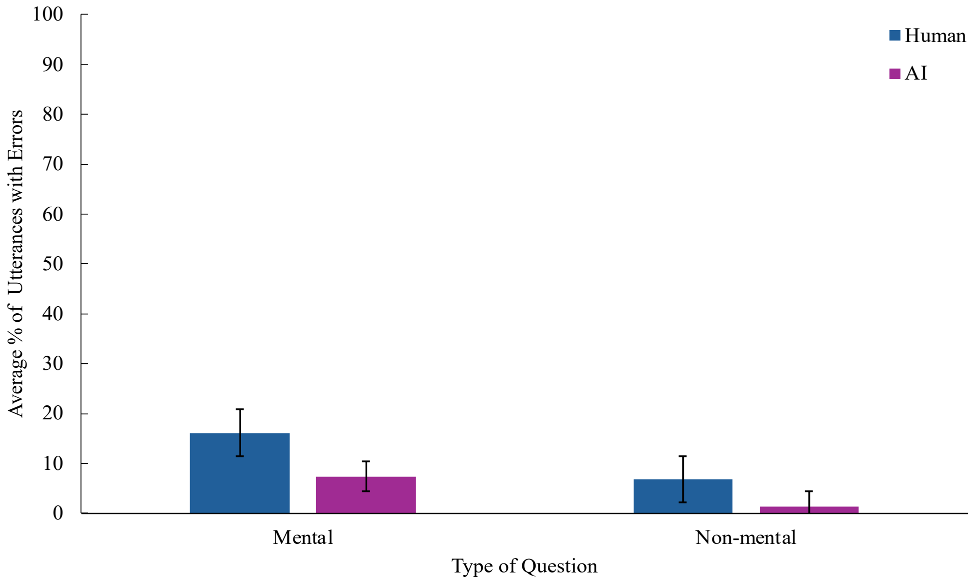
Comparison of Topical Relevance



Note. (A) Average Percentage of Maze Words & (B) Average Number of Filler Pause Words. Figures 7A displays the percentage of maze words that form participants' ($N = 60$) overall responses relative to the type of question asked by the human or AI conversational partner. Figure 7B shows the mean number of filler pause words spoken, which are a specific type of maze word. Error bars represent standard error.

Figure 8

Comparison of the Percentage of Utterances with Errors



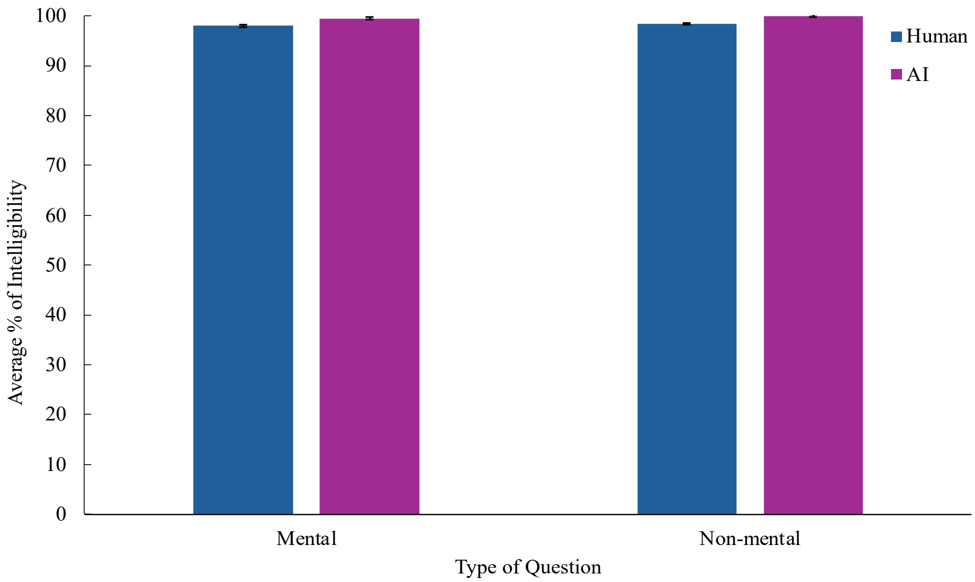
Note. Figure 8 shows the average percentage of utterance-level errors produced by participants (*N* = 60) within a sample relative to the conversational partner and type of question asked. Error bars represent standard error.

Intelligibility

Finally, a main effect for the type of conversational partner on the participants' intelligibility was observed. While children generally maintained a high level of intelligibility, with the average score always hovering above 98%, children responded to the AI voice agent with significantly more intelligible responses compared to children who were paired with the human conversational partner, as shown by Figure 9. However, no significant difference in intelligibility was observed when considering the type of question asked.

Figure 9

Comparison of Intelligibility



Note. Figure 9 displays the average percentage of overall intelligibility calculated for participants (*N* = 60) based on the type of question asked by the human or AI conversational partner. Error bars represent standard error.

Discussion

Variations in Story Comprehension

The main purpose of this study was to examine the efficacy of story-listening with AI as compared to with a human. I hypothesized that children would overall be equally accurate in answering story-related questions when conversing with both a human and AI. This prediction was partially supported by the data. The results of the post-task assessment illustrate that the AI voice agent successfully scaffolded a comparable level of story comprehension in children when comparing the performance of children paired with the human counterpart (see Table 1). However, the observed interaction between the AI voice agent and the type of question asked revealed lower story comprehension in children who answered non-mental questions asked by the AI voice agent.

In comparison to the post-task, unexpected differences in performance within participants during the story-listening task were observed. Children paired with the AI voice agent exhibited decreased story comprehension compared to children paired with the human conversational partner. Additionally, children responded to non-mental questions with greater accuracy, which diverged from the interaction detected in the post-task, linking lower story comprehension to non-mental questions asked by the AI voice agent. Since only the in-task assessment demonstrated a main effect for the type of question asked, it is difficult to establish a clear trend outlining its relationship to children's understanding of the story. The decrease in story comprehension observed in participants assigned to the AI condition, during the story-listening task, suggests that the AI voice agent is not yet as effective as a human in scaffolding story comprehension in real time. Children in both groups appeared to equally understand the story based on their similar performance on the post-task. The difference in administration between the in-task and the post-task, which was given after the story ended, as well as the participants' performance, suggests that AI might lack effectiveness in prompting strong interactions in real time, but that it *can* be an effective tool for story comprehension depending on the context.

Similarities in Verbal Responses

It was also initially predicted that children would exhibit greater verbal engagement with the human conversational partner compared to with the AI voice agent. This prediction was not fully supported, as more similarities were observed than expected between the human and AI voice agent. While some trends depicted better communication with a human, others failed to indicate a significant difference between the two assigned conversational partners. Namely, there were no significant differences between the human and AI voice agent in terms of lexical diversity, and several subcomponents of language productivity, such as MLU, the percentage of abandoned utterances, and the percent of time spent paused before participants answered a question. On the other hand, participants spoke faster, produced more words overall, more filler pause words, and more errors in response to the human conversational partner.

Interestingly, not all the trends correlated with the human conversational partner are indicative of more effective verbal engagement. While children responded to the human conversational partner with longer responses overall, as seen by a significantly higher average NTW, their responses also contained a greater percentage of utterances with errors, more filler pause words, and a higher percentage of maze words. These three trends are closely linked, as filler pause words are a specific type of maze word and utterance-level errors commonly include maze words that detract from the overall meaning and syntax of the utterance. The increased presence of errors highlights that a longer response does not necessarily equate to a more accurate response, emphasizing that the general decrease in NTW observed in the AI condition is not an indicator of AI being an ineffective tool, especially since MLU, the established metric for syntactical complexity and language development, was comparable between the two conditions. Instead, the differences in these trends highlight the capability of AI in establishing equally complex, and perhaps even more accurate, responses as compared to a human.

Behavioral Differences

Despite prompting accurate responses, differences in the participants' behavior were observed during the testing sessions. Examiners frequently noted that children were hesitant to tell the AI voice agent when they were unsure of the correct response during the story-listening task. Children often had to be prompted by a human examiner to alert the AI voice agent. This trend was partially reflected in the data, as a marginally significant trend was observed in terms of pause time percentage; on average, children took longer to answer a question asked by the AI voice agent. A similar finding was reported by Gampe et al., (2023), who found that children were less likely to attempt to communicate with the AI voice agent when they were unsure of what to do. This trend might be explained by the way that children view AI. Because children do not fully view AI as human, it is likely that they might not feel the need to communicate their thought process as thoroughly as they would with a person (Gampe et al., 2023; Lee & Jeon., 2022).

The decrease in filler pause words observed in the AI condition mirrors children's decreased attempts to maintain communication, as filler pause words are a verbal cue often used to communicate that the speaker is still expressing a thought (Watanabe et al., 2008). Given that a major function of mainstream AI voice agents, like Siri and Alexa, is to answer our questions and provide information, it is possible that children pre-exposed to smart devices might think that the seemingly omniscient AI voice-agent in the study is able to pick up on their thoughts without their input, suggesting that theory of mind in the context of AI might develop differently. Further research is necessary to properly explore the implications of this trend in terms of how children view AI, as children have been found to anthropomorphize AI to a greater extent than humans (Lee & Jeon., 2022)

At this stage, it appears that the initial prediction that verbal engagement would be lower with the AI voice agent was not supported, as the AI voice agent was surprisingly effective at eliciting similarly complex responses that were equally as lexically diverse as that of children paired with a human. Additionally, children responded to the AI voice agent with more syntactically accurate phrases that contained less utterance-level errors, filler pause words and maze words when compared to children in the human condition. In sum, the AI voice agent was more effective at prompting narratively rich responses than previously expected – it is almost as effective as its human counterpart. Despite delivering promising results, the decreased story comprehension and observed challenges in communication between the participants and the AI voice agent indicate that future changes to the AI voice agent's script and tone are needed to facilitate greater verbal interactions and by extension greater verbal engagement to adequately match that of a human.

Limitations

The straightforward question-answer format of the story-listening task is another factor that could have potentially limited language productivity. It is possible that the interview question-answer style might have prevented participants from fully expressing their thoughts as they would in a more natural conversation setting, inaccurately representing their full language capability. Additionally, while all participants responded to the same 20 questions, with the same fact-type format, to establish an accurate comparison between the groups, the lack of other types of questions commonly asked during dialogic reading might decrease the external validity of this study. It is possible that children might respond differently to the AI voice agent if asked open-ended or multiple-choice questions, as different types of questions have been shown to vary in terms of accuracy and the way they are comprehended (Ozuru et al., 2013). A follow-up study examining additional questions might provide greater insight into this topic.

Another potential limitation of the design is that the lengthy testing sessions, which ranged from 3 to 4 hours, could have fatigued the children, and artificially decreased their performance. To mitigate this, most participants underwent the story-listening task and completed both story comprehension assessments first. Additionally, participants were encouraged to take breaks and given a standard 5-minute break in between the story listening task and behavioral assessments.

Conclusion

Despite increasing story comprehension by developing emergent literacy, dialogic reading is often underused by parents and teachers due to logistical constraints like time. For example, it would be nearly impossible for one instructor to engage in one-on-one dialogic reading with their students while still maintaining classroom order. Our investigation aims to bridge this gap by demonstrating the potential efficacy of AI as a tool for learning, specifically through dialogic listening, a similar activity to dialogic reading that scaffolds story listening comprehension. The results of this study demonstrate that AI voice agents are almost as effective as humans, and the exponential rate of technological advancements are likely to continue narrowing this gap in the near future. Thus, our research underscores the importance of designing more accessible AI systems tailored for educational purposes so that all children may receive quality education in the future, regardless of their living situation.

The results of this study raise the broader question of how children exposed to multiple languages differ in their interactions with AI. Approximately one in five students in the American school system come from immigrant families, with many of them speaking a second or third language, making it important to consider the effect of language and culture as well (Fix & Passel., 2003). The ultimate goal of this study is to bring awareness to the limitless potential of AI as a learning tool in a bid for educators to eventually implement it into their curriculum. For this reason, we have begun incorporating bilingual Chinese and Spanish-English children to investigate how language influences children's interactions with AI across diverse populations in order to provide a more representative evaluation of AI as a tool for learning.

References

Allen, M.L. (2013). Kaufman assessment battery for children, Second Edition. In: Volkmar, F.R. (eds) Encyclopedia of Autism Spectrum Disorders. Springer, New York, NY. Doi:10.1007/978-1-4419-1698-3_94

Beaudoin, C., Leblanc, É., Gagner, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, *10*, 2905. Doi: 10.3389/fpsyg.2019.02905

Buchanan, B. G. (2005, November/December). A (very) brief history of artificial intelligence. *AI Magazine*, *26*(4), 53. Doi: 10.1609/aimag.v26i4.1848

Fix, M., & Passel, J. S. (2003). U.S. immigration – Trends & implications for schools. New Orleans; The Urban Institute. Retrieved from https://webarchive.urban.org/UploadedPDF/410654_NABEPresentation.pdf

Fergadiotis, G., Wright, H. H., & West, T. M. (2013). Measuring lexical diversity in narrative discourse of people with aphasia. *American Journal of Speech-Language Pathology*, *22*(2), S397–S408. DOI: 10.1044/1058-0360(2013/12-0083)

Gampe, A., Zahner-Ritter, K., Müller, J., & Schmid, S (2023). How children speak with their voice assistant sila depends on what they think about her. *Computers in Human Behavior*, *143*, 2023, 107693, Doi: 10.1016/j.chb.2023.107693

Kaufman, A. S., & Kaufman, N. L. (2004a). *Kaufman assessment battery for children: Second edition (KABC-II)*. Circle Pines, MN: American Guidance Service.

Kleeck, A., & Whitehurst, G. (2009). Dialogic reading: A shared picture book reading intervention for preschoolers. *On Reading Books to Children: Parents and Teachers*, Routledge, New York, 2009, 170-173.

Lee, S., & Jaeho J. (2022) Visualizing a disembodied agent: Young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning*, 1–26, Doi: 10.1080/09588221.2022.2067182

Marks, R. A., Labotka, D., Sun, X., Nickerson, N., Zhang, K., Eggleston, R. L., Yu, C. L., Hoeft, F., Uchikoshi, Y., & Kovelman, I. (2022). Morphological awareness and its role in early word reading in English monolinguals, Spanish-English, and Chinese-English simultaneous bilinguals. *Bilingualism: Language and Cognition*, *26*, 268-283. Doi:10.31234/osf.io/xpycj

Miller, J & Iglesias, A. (2020). Systematic Analysis of Language Transcripts (SALT), Version 20 [Computer Software]. Madison, WI: SALT Software, LLC.

Okon, J. J. (2011) Role of non-verbal communication in education. *Mediterranean Journal of Social Sciences*, *2*(5). DOI: 10.36941/mjss

OpenAI. (2023). Whisper AI (September 14 version) [Large language model]. https://colab.research.google.com/drive/1pisVdIIjaYSve_FH0mWXOV-Xz53kl7Qs

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale, 67*(3), 215–227. Doi: 10.1037/a0032918

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). Woodcock-Johnson IV tests of achievement. *Sage Journals*, *33*(9), 391-398, Doi :10.1177/0734282915569447

Strouse, G., Nyhout, A., & Ganeaá, P. (2018). The role of book features in young children's transfer of information from picture books to real-world contexts. *Frontiers in Psychology*, *9*, 2018, Doi: 10.3389fpsyg.2018.00050

Valle, A., Massaro, D., Castelli, I., & Marchetti, A. (2015). Theory of mind development in adolescence and early adulthood: The growing complexity of recursive thinking ability. *Europe's Journal of Psychology*, *11*(1), 112–124. Doi: 10.5964/ejop.v11i1.829

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A. (1999). Comprehensive test of phonological processing: CTOPP. Austin, TX: Pro-ed.

Watanabe, M., Hirose, K., Den, Y. & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81-94, Doi: 10.1016/j.specom.2007.06.002

Whitehurst, G., & Lonigan, C. (1998). Child development and emergent literacy. *Child Development*, *69*(3), 848–872, Doi: 10.1111/j.1467-8624.1998.tb06247.x

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers &amp; Education*, *161*, 104059, Doi: 10.1016/j.compedu.2020.104059

Xu, Y., Aubele J., Vigil, V., Bustamante, A., Young-Suk, K., & Warschauer, M. (2021). Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement. *Child Development*, *93*(2), Doi: 10.1111/cdev.13708

Zaiontz, C. (2023). Real Statistics Resource Pack software release 8.9.1 [Microsoft Excel] Retrieved February 15, 2024, from www.real-statistics.com

Appendix A

Additional Comparison of Language Productivity between the Human and AI Conditions

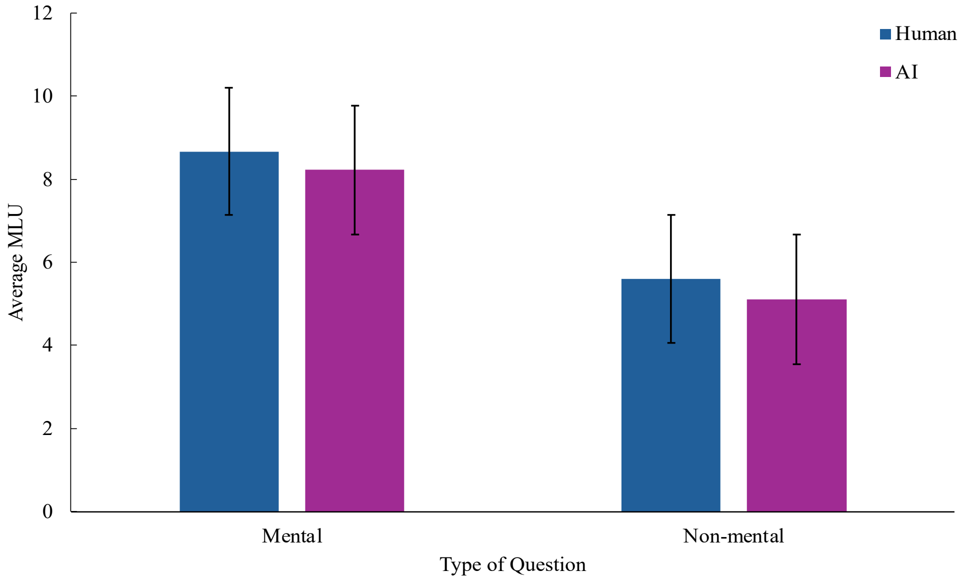


Figure A1. The average MLU between the human and AI conditions based on the type of question asked.

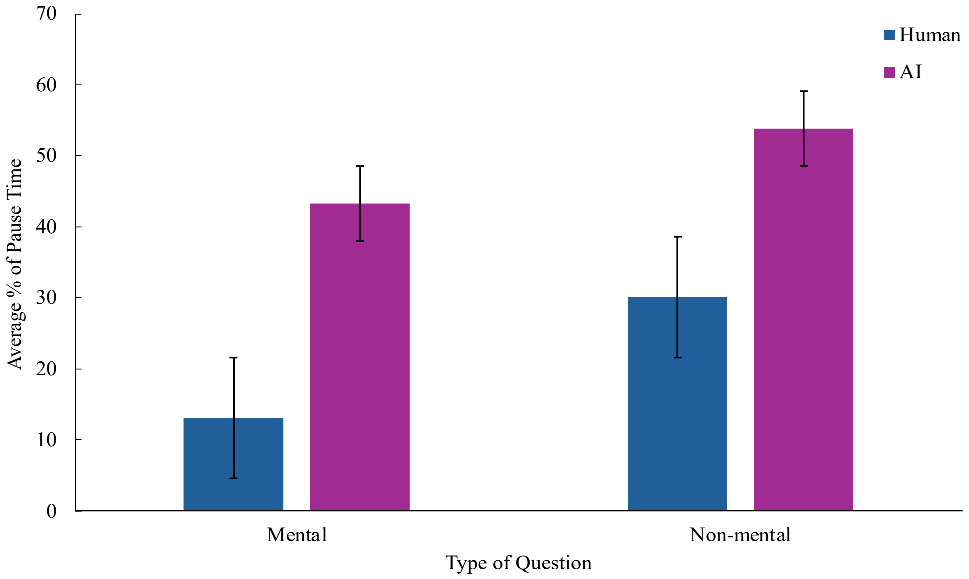


Figure A2. The average percentage of time spent paused between the human and AI conditions in the context of the type of question asked.

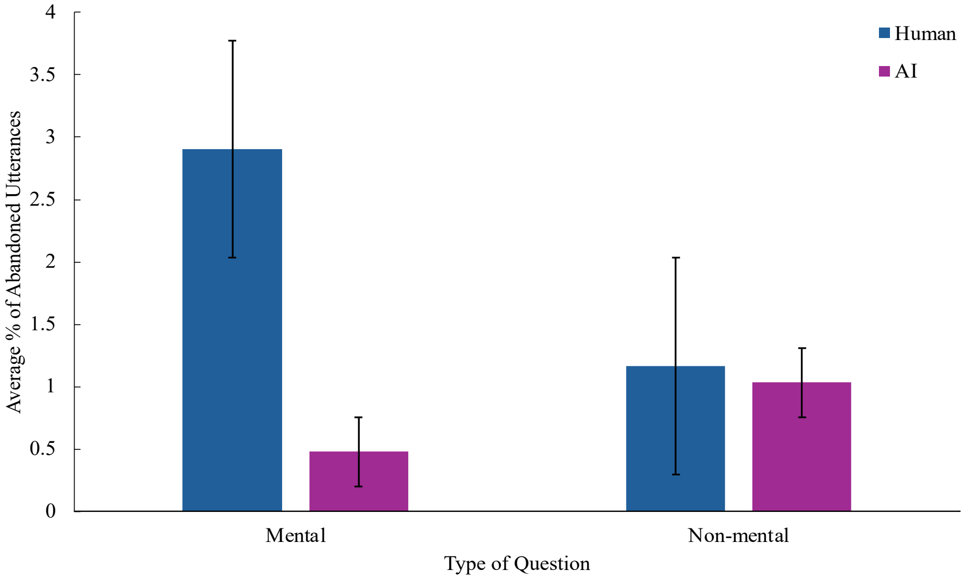


Figure A3. The average percentage of abandoned utterances between the human and AI conditions in terms of the type of question asked.