AI doesn't produce child-directed speech like caregivers do

Esther Werbach (ewerbach@sas.upenn.edu)[1], Margaret Kandel[2], Annika Heuser[3], Kathryn Schuler[3]
[1]Department of Cognitive Science, University of Pennsylvania
[2]MindCORE, University of Pennsylvania
[3]Department of Linguistics, University of Pennsylvania

Abstract

AI-generated child-directed speech is increasingly positioned as a supplement or replacement for caregiver speech, either for use in early-learning technologies (AI-powered toys) or as a substitute to human data for research. We compare child-directed speech from AI and human caregivers along dimensions linked to child language outcomes: utterance length, lexical diversity/complexity, pronoun use, wh-words, and verb types. While the AI-generated speech is longer than caregiver input, it is less variable and exposes children to fewer word types. AI and caregiver speech also show different distributions of grammatical elements: AI favors first- and second-person pronouns, while caregivers use a broader set of referential forms; AI produces primarily argument wh-words, whereas caregivers use wh-words that appear in more complex syntactic constructions; and AI tends to use different kinds of verbs. These findings suggest that AI may not sufficiently imitate child-directed speech with the properties required for early language learning.

Keywords: child-directed speech; language acquisition; artificial intelligence; large language models

Introduction

Young children's exposure to artificial intelligence (AI) is rapidly increasing (Mann et al., 2025). As AI-powered toys such as Curio, Miko, and FoloToy are marketed as educational tools and screen-time alternatives, AI systems are likely to become regular conversational partners for young children. While these toys have been pitched as tools for enhancing young children's cognitive, social, and linguistic development, empirical evidence for these claims is lacking (Xiao & Gonçalves, 2025), and it remains unclear whether interactions with AI systems are able to provide the proper scaffolding for healthy development. This uncertainty leaves developmental scientists concerned about the potential risks of expanding children's exposure to AI and replacing human caregiver interactions with AI systems (Roche et al., 2025).

One area of development that is highly dependent upon interaction with caregivers is language acquisition. In order for young children to become successful users of a language, they must not only learn the words of their language (vocabulary development) but also how these words are structured and modified (morphological development) and how they combine to form sentences (syntactic development). Developing this abstract system of rules (i.e., the grammar of their language) requires linguistic input that is abundant, personalized, and syntactically diverse (e.g., Huttenlocher et al., 2002). While the large language models (LLMs) that AI systems are built upon effectively mimic natural language, LLM output is often more homogeneous than language produced by humans (Zanotto & Aroyehun, 2024), which could mean that AI-generated child-directed speech may not display the lexical and syntactic variability required for child language learning.

Moreover, given recent claims that LLMs may be able to replace humans in behavioral research (Argyle et al. 2023; Dillion et al., 2023; Meng, 2024), it is crucial to understand how human- and AI-generated child-directed speech differ. Corpora of child-directed speech are difficult to compile, so researchers may be motivated to supplement existing corpora with AI-generated child-directed speech. However, determining the specific aspects of child-directed speech that contribute to language learning is still an active area of investigation. We do not yet know what input is necessary and sufficient for language learning, and if AI-generated child-directed speech differs from that of human caregivers, then studying LLM output is unlikely to provide the answer to this question.

These issues thus motivate our central research question: Does AI-generated child-directed speech resemble caregiver input along the dimensions known to support early language learning?

Input Characteristics that Support Language Learning

Prior work on child language acquisition has identified several features of child-directed speech that influence children's language learning outcomes. Here, we introduce a subset of these learning-relevant features that we will focus on in the present study.

Child- and child-directed speech is often characterized in the language acquisition literature by properties such as utterance length (tokens), lexical diversity (types), and vocabulary complexity. These properties are typically operationalized as mean length of utterance (MLU), standardized type-token ratio (STTR), and estimated vocabulary age of acquisition (AoA), respectively. Here, we use these metrics to characterize child-directed speech in both human caregivers and AI.

Beyond high-level measures like type and token frequency, differences between caregiver and AI child-directed speech may also appear in the *kinds* of

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

linguistic elements they use. Because children learn how grammatical elements are used from the distributional structure of the input, differences in the composition of child-directed speech may have consequences for learning, even when overall measures appear similar. To examine whether AI-generated child-directed speech differs from caregiver child-directed speech in these ways, we focus here on three classes of elements that are both frequent in child-directed speech and central to early language learning: pronouns, wh-words, and verbs.

Pronouns are one commonly studied aspect of children's language acquisition. Pronouns require children to track how forms map onto discourse roles rather than stable referents (e.g. 'he' does not always refer to the same individual). In many languages, pronouns additionally encode syntactic information, such as case (e.g., subject vs. object forms), which in turn supports the identification of grammatical roles and argument structure. This is particularly important in languages such as English, where overt case marking is limited. Pronouns furthermore have been shown to provide an anchor point for generalizing syntactic structures (e.g., Childers & Tomasello, 2001), making pronoun exposure an important entry point into grammar learning.

Wh-questions are another commonly studied aspect of children's language acquisition. They provide cues about how grammatical structure is linked to the discourse and serve as examples of complex syntactic structures that children must learn (e.g., long-distance filler-gap dependencies, syntactic movement). Exposure to these constructions can provide cues to underlying hierarchical relations, particularly in cases of syntactic movement, where the displaced wh-elements are interpreted in relation to their original argument position. Since different types of wh-questions vary in structural complexity (e.g., argument vs. those involving movement across clauses), their distribution in the input provides insight into the range of syntactic constructions children are exposed to.

Finally, verbs encode core relational and event-structural information that is central to children's development of argument structure. Moreover, the verbs used can also provide information about the type of information discussed in a discourse. In this study, we categorize verbs into four semantic classes that capture meaningful distinctions in verb meaning and usage (Levin, 1993). Since AI systems (in particular, LLMs) often lack access to a shared perceptual environment with the child, it is possible that the distribution of verb types within AI-generated child-directed speech may differ from that of human caregivers in meaningful ways, reflecting differences in the discourse content and argument structures that children would be exposed to.

Characteristics of AI-generated Language

To date, there has been no systematic examination of whether AI-generated child-directed speech captures the learning-relevant properties of naturalistic caregiver input, including its variability, distributional structure, and use of grammatical elements. However, existing work already suggests that the linguistic profiles of LLM-based AI systems differ from those of humans. For example, a comparison of human-written and LLM-generated texts across 250 linguistic features found that human texts exhibit substantially greater variability than LLM outputs (Zanotto & Aroyehun, 2024). A recent benchmarking study investigating how well LLMs (GPT-4o and Llama 3) can approximate child and caregiver language at the word and utterance levels found that they consistently fall short on dialogue-level dynamics, tending to over-align with the interlocutor and produce less diverse responses than human caregivers, especially in multi-turn interactions (Liu & Fourtassi, 2024). These findings raise the possibility that AI-generated child-directed speech could be less variable than human caregiver speech, potentially limiting children's exposure to linguistic structures important for acquisition.

The present study builds upon these findings to explore more specifically how AI-generated child-directed speech compares to naturalistic caregiver input along learning-relevant properties. In this study, we investigate caregiver child-directed speech alongside AI-generated child-directed speech to determine if AI chatbots produce child-directed input that exhibits the same structural and distributional properties that promote language learning in human caregiver speech.

Methods

We compare samples of child-directed speech produced by human caregivers and AI in response to 3 year-old children's utterances from the CHILDES corpus (MacWhinney, 2014). We focused on this age range because 3-year-olds are still in a period of active language development, but they have acquired enough language to participate in conversation. This age range also corresponds to the youngest (and therefore most vulnerable) of the target demographic for AI toys, making it an appropriate window for comparing caregiver and AI-generated input.

Caregiver Child-Directed Speech Sample

Our sample of caregiver child-directed speech consisted of caregiver responses taken from the Providence corpus in CHILDES (Demuth et al., 2006; MacWhinney, 2014). In this corpus, North American English speaking children were recorded every few weeks while engaged in natural conversation with their caregivers (Demuth et al., 2006). We focus on the recordings of the five of these children and their mothers who were recorded during the 3-year-old range (see Table 1). For each child, we extracted the transcriptions of all utterances produced by the mother and the child during this age window. We used the mothers' utterances as our caregiver child-directed speech sample and used the child utterances to generate AI child-directed speech (see below).

Table 1: Characteristics of 3-year-old children included in the analysis (Providence corpus, CHILDES).

Child name	Age range (months)	Mean age (months)	Utterances (n)	Sex
Alex	36.2 – 41.5	39.1	7868	M
Lily	36.1 – 46.8	40.9	8114	F
Naima	36.0 – 46.3	40.3	3402	F
Violet	37.6 – 47.8	42.5	3026	F
William	36.4 – 40.6	38.4	5410	M

AI-generated Child-Directed Speech Sample

Our sample of AI-generated child-directed speech was produced by the conversational language model ChatGPT-5 mini (OpenAI, 2026). We selected this model because OpenAI's ChatGPT models are among the most popular AI chatbots (Bickham et al., 2024; Chatterji et al., 2025) as well as among those integrated into AI-powered toys for children (e.g., Curio; see Pavliv et al., 2025).

We use the *kani* package (Zhu et al., 2023) to prompt ChatGPT-5 mini to respond to each child utterance in the selected CHILDES transcripts. This allowed us to construct a sample of AI-generated child-directed speech matched to the same child utterances used in our caregiver child-directed speech sample. We provided a fixed system prompt that instructed the model to adopt the role of a parent interacting with their child:

> *You are the parent of a 3-year-old English-speaking child.\n Your child's name is [NAME].\n You are having a conversation with your child.\n Your goal is to maintain the interaction and support the child's attempt to communicate.\n yyy and xxx indicate unintelligible speech.\n*

The task prompt then specified how the model should respond to each child utterance and enforced response constraints to promote consistency across model responses to different child transcripts. Following prior work, we limited model response lengths to 50 tokens (Liu & Fourtassi, 2024). The prompt used was as follows:

> *For each child utterance below, respond naturally as the parent.\n Ensure your response is no longer than 50 words regardless of the prompt.\n Do not include any commas or line breaks in responses.\n Return your answer in CSV format with this header exactly:\n utterance_id, response\n Do not add any other text.\n*

For each child utterance in the selected transcripts, the model generated a single response under these constraints. Child utterances were submitted sequentially in batches of 50, which both reduced the number of API calls and also gave the model access to some of the conversational context (an approximation of the discourse history available to the children's mothers).

Data Processing

Prior to analysis, we normalized the caregiver and AI-generated speech samples by removing punctuation and converting the text to lowercase. We then used *UDpipe* (Wijffels et al., 2019) to tokenize, lemmatize, and part-of-speech (POS) tag the data. We excluded any unintelligible tokens (*xxx*, *yyy*)[1] and other elements flagged as non-linguistic by *UDpipe*'s POS tagger (e.g. SYM); this resulted in exclusion of < 1% of the tokens in the data.

Analysis

Characterizing Input Complexity

We first characterized AI and caregiver child-directed speech with a set of commonly used metrics from the language acquisition literature: mean length of utterance (MLU), standardized type-token ratio (STTR), and vocabulary age of acquisition (AoA) estimates. These metrics are widely used to index properties of child- and child-directed speech, allowing us to situate our results within the existing empirical literature. All metrics were computed separately for the caregiver and AI-generated speech samples using a shared analysis pipeline.

MLU Mean length of utterance is widely used as a measure of linguistic maturation for language acquisition research and serves as a general index of utterance-level complexity (e.g., De Villiers & De Villiers, 1973). Larger MLUs correspond to more language input for children and are associated with improved vocabulary learning outcomes (e.g., Bang & Nadig, 2015) and greater opportunity for learning more complex syntactic structures (more complex syntactic input is associated with greater syntactic growth; e.g., Huttenlocher et al., 2002). We calculated MLU by taking the average number of word tokens across all utterances by child for both speech sources (caregiver, AI).

STTR Standardized type-token ratio is used to quantify lexical diversity in child-directed speech, controlling for utterance length (e.g., Richards, 1987). Lexical diversity is important not only for children's vocabulary development (the more words children are exposed to, the more words they have the opportunity to learn), but also for children's acquisition of the morphological rules of their language. Exposure to more types in the input allows children to better learn generalizations across words (Richtsmeier et al., 2011; Yang, 2016). For each child and source (caregiver, AI), we segmented the data into consecutive 50-word token windows; then, we calculated the type-token ratios within each window and took the average.

[1] These tokens occurred either when the caregiver response was not intelligible or when the AI-generated responses used these tokens to refer to something unintelligible that a child had said (e.g., 'what is this yyy show me please').

AoA Vocabulary age of acquisition refers to the age at which a word is typically learned. This metric, commonly used in studies of early word learning, serves as a proxy of lexical complexity within child-directed speech, providing insight into how closely the input aligns with the child's stage of linguistic development (Portelance et al., 2020). The use of more advanced vocabulary within a child's input increases the potential learning opportunities for the child. We used *wordbankr* (Frank et al., 2017) to estimate the age of acquisition for all words available in the database. Wordbankr aggregates data from the MacArthur-Bates Communicative Development Inventory (MB-CDI; Fenson et al., 2006). Age of acquisition was defined as the age at which 50% of children were reported to produce the word. Using these data, we computed a frequency-weighted mean age of acquisition for each child and source (AI and caregiver).[2]

Characterizing Grammatical Form Usage

Children's input must contain enough syntactic variability for children to learn the grammatical rules of their language. We focus on three features of the input that may affect children's grammar learning: pronoun use (which provides cues to the English case system), the presence of wh-words (which are associated with complex syntactic structures such as long-distance dependencies), and verbs (which encode how actions and perceptual experiences are structured in input).

For each child, we computed the proportion of tokens accounted for by each type (pronoun types, wh-word types, and verb types) separately for caregiver and AI responses and then calculated the difference between the two. For verbs, we focused our analysis on four verb type categories: cognition, communication, motion, and perception. We selected 4 representative verbs for each category (Levin, 1993): cognition (*mean*, *understand*, *think*, *know*), communication (*say*, *tell*, *ask*, *talk*), motion (*come*, *put*, *get*, *go*), and perception (*hear*, *listen*, *look*, *see*).

Results

Overview

We begin with an overview of our three primary language acquisition metrics. Figure 1 shows a boxplot of each metric computed for both AI-generated and caregiver child-directed speech.

Overall, we find that AI uses more tokens in its child-directed speech, but caregivers use more types. Mean length of utterance (MLU) was higher in AI-generated child-directed speech than in caregiver child-directed speech (*paired t-test*: $t(4) = 11.83$, $p < .001$, 95% CI [3.57, 5.76]). But the standardized type-token ratio (STTR) was higher for caregivers than for AI (*paired t-test*: $t(4) = -3.25$ $p = .031$, 95% CI [-0.07, -0.006]). In the sections that follow, we unpack each of these findings in more detail.

For age of acquisition (AoA), although the visualization in Figure 1 suggests a possible trend towards more complex vocabulary in caregiver speech, we find no significant difference in frequency-weighted mean age of acquisition (AoA) between AI-generated and caregiver child-directed speech (*paired t-test*: $t(4) = -1.81$, $p = .14$, 95% CI [-0.37, 0.08]).[3]

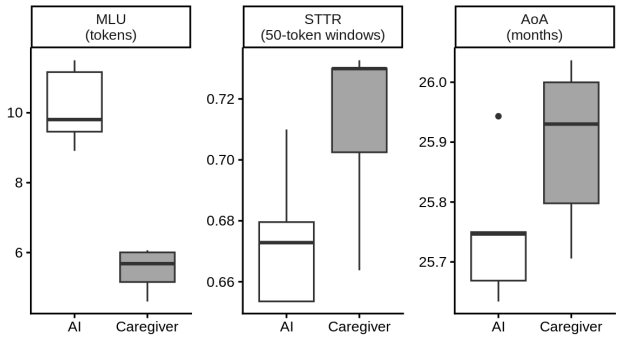

Figure 1: Language acquisition metrics in caregiver and AI-generated responses. Boxplots show medians and interquartile ranges.

AI Uses More Tokens than Caregivers

As shown above, our AI-generated child-directed speech uses significantly more tokens per utterance than the children's caregivers, even though the prompt included a 50-word response limit. However, although the AI produced longer utterances on average (10.17 tokens per utterance) than caregivers (5.50 tokens), the distributions differ dramatically: AI utterance lengths are approximately symmetric, but caregiver utterances are heavily skewed, with many single-token responses and a long tail of longer utterances (see Figure 2).

Examining these single-token utterances shows that they mainly consist of brief acknowledgements (e.g. 'okay', 'yeah', 'oh', 'no', yes'), while concrete nouns (e.g. 'dog', 'cat', 'ball') appear primarily consistent with ostensive labeling (Axelsson et al., 2012).

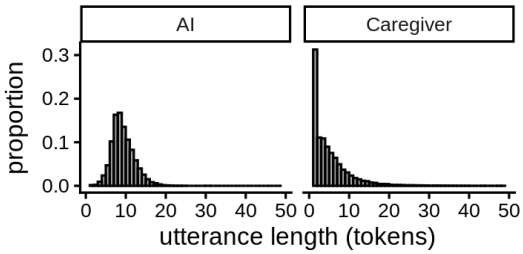

Figure 2: Utterance length distribution for caregiver and AI-generated responses.

[2] We also computed this analysis using age of acquisition ratings from Kuperman et al. (2012) and observed the same pattern of results.

[3] See deWinter (2013) for discussion of the feasibility of conducting *t*-tests with small sample sizes.

Caregivers Use More Types than AI

In our sample, caregivers and AI also differed in the number of word types they produced. Across all children, caregivers produced on average 2389 different word types (SD = 1001), whereas the corresponding AI responses included on average only 1826 different word types (SD = 336). Figure 3 shows the number of word types used by each child's caregiver and the AI as a function of the number of tokens. We find that all five children encounter many more word types in the child-directed speech of their caregivers compared to our AI-generated child-directed speech; this difference is also captured in our STTR measure (Mean caregiver STTR = 0.71, SD = 0.03; Mean AI STTR = 0.67, SD = 0.02). In the most dramatic example, by the time Lily's mother has uttered 20,000 tokens, she has exposed Lily to over 2000 different word types. However, in the same number of tokens, our AI-generated child-directed speech would have exposed Lily to fewer than 1000 different types. This difference has potential implications for learning, as exposure to more diverse word types is necessary for learning grammatical generalizations across words (Yang, 2016).

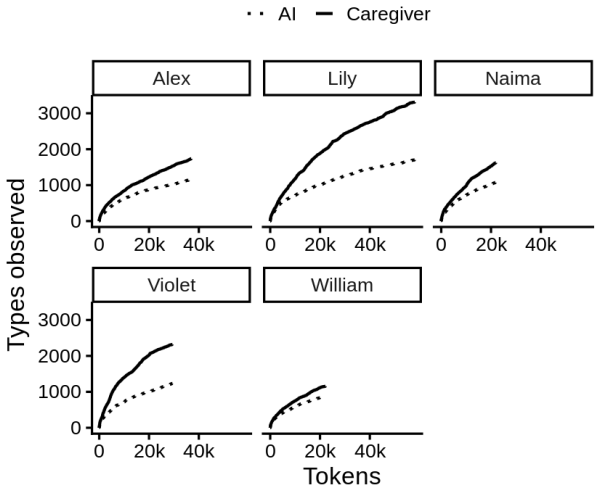

Figure 3: Word types observed in caregiver (solid) and AI (dashed) responses as a function of tokens for each child. AI token counts are capped to caregiver counts for comparison.

AI and Caregivers use Pronouns, Wh-words, and Verbs Differently

Beyond simply asking whether AI responses differ from caregivers in their type and token count, we can also ask whether patterns of use of particular words differ between caregivers and AI. In Figure 4, we show that AI and caregivers differ in their patterns of pronoun use. AI uses 'you', and 'me' more when responding to 3-year-olds, while the children's caregivers use more 'we', 'she', 'they', 'he', and 'i'. This pattern may reflect AI's lack of access to a shared social and physical environment, limiting its reference to the immediate conversational partner (the child), while caregivers naturally draw on broader social context.

Figure 4 shows a similar pattern for wh-words. The AI model is more likely than caregivers to ask children 'what', 'where', and 'which', while caregivers are more likely than AI to ask children 'how', 'who', and 'when'. This suggests that caregivers produce more adjunct wh-questions (e.g., *how*, *when*) than AI, exposing children to question forms that children do not learn to consistently invert as quickly as argument wh-questions (e.g. *what*, *which*) (Pozzan & Valian, 2017). This demonstrates that caregiver input may provide more opportunities to learn adjunct wh-questions.

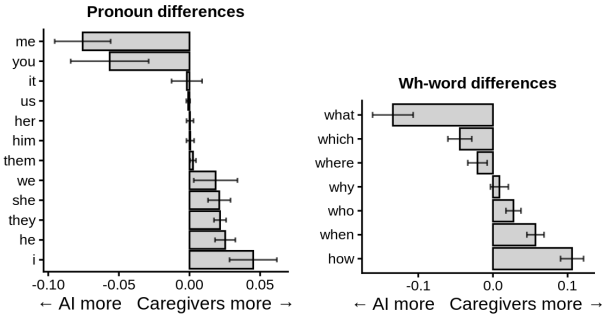

Figure 4: Differences in the use of pronouns and wh-words between caregiver and AI-generated child-directed speech. Bars show mean normalized frequency differences (caregiver - AI).

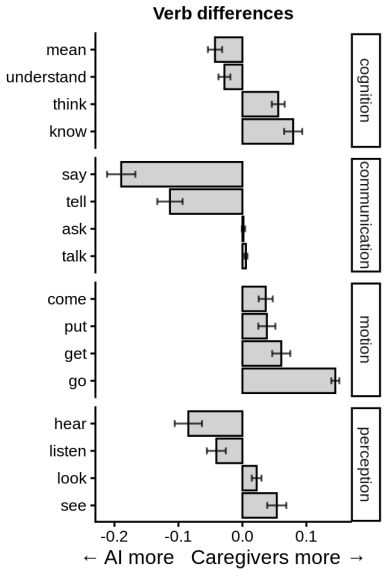

Figure 5: Verb differences for caregiver and AI-generated responses.

Finally, we asked whether we observe any differences in the way AI and caregivers use verbs (Figure 5). As a reminder, we focused our analysis on four categories of verbs: cognition, communication, motion, and perception. AI produced more communication verbs, while caregivers' speech contained more motion verbs. Cognition verbs were

split among AI and caregiver speech, with AI input containing more *mean/understand* and caregiver input containing more *think/know* (epistemic state verbs). One possible explanation for these results could come from the differing goals of AI systems and caregivers, as caregivers' motivation may align more with sharing their thoughts and experiences with their child.[4] We also found a split amongst perception verbs: AI uses *hear/listen* more while caregivers use *look/see* more. Since *look/see* refer to perceptual experiences that are jointly available in the immediate environment, caregivers may use these verbs more frequently to guide joint attention between themselves and the child.

General Discussion

Overall, our findings indicate that child-directed speech produced by AI differs in profile and composition from that produced by human caregivers. Using common speech characterization metrics from the language acquisition literature, we find that while AI speech contains longer utterances with more tokens, caregiver speech contains a more varied distribution of utterance length and exposes children to more word types. Furthermore, the AI and caregiver speech also differ in their usage of pronouns, wh-words, and verbs, with caregivers' speech drawing on more shared perceptual, experiential, and social context than AI. Prior literature suggests that these differences may have negative implications for young children's language development. In particular, AI-generated speech appears to provide less evidence of word and sentence types participating in morphological and syntactic rules (though note that more research is needed to determine precisely which aspects of children's language input are crucial for grammar learning; Kempe et al., 2024).

This study constitutes an initial step toward investigating the differences between child-directed speech produced by humans and AI, with the goal of understanding the potential impacts of young children's increased exposure to AI. A current limitation of the present work is that we compare a single AI model to speech from five different caregivers, each of whom naturally varies in their linguistic input. In future work, we plan to compare the output of multiple LLMs (e.g. Gemini, Claude) and model sizes to quantify the variability of different models vs. that of different caregivers. Additionally, some of the differences that we observed between the caregivers' and Chat GPT-5 mini's responses could plausibly be reduced through more targeted prompting. For example, explicitly permitting or encouraging brief acknowledgements (e.g. 'okay', 'yeah') may lead the model to generate them more frequently, better matching the utterance length distribution of the caregivers.

Furthermore, in the current study, we focus solely on properties of the child's linguistic input, however early language development depends not only on input properties but also on socio-cognitive factors such as social contingency, joint attention, and the motivation to communicate (e.g., Bruner, 1978; Snow, 1977). Future work will additionally need to address the role of these factors in AI–child interactions to determine whether they provide the necessary pedagogical scaffolding for language development. Our verb analysis provides some initial evidence for the role of social context in shaping AI interactions, from the perspective of linguistic input. The higher frequency of motion and perception verbs in caregiver speech (e.g. *come*, *put*; *look*, *see*) could reflect the fact that caregiver speech inherently involves real-time coordination, in which the caregiver and the child jointly attend to objects and actions in the same 3-dimensional space; in contrast, AI-generated speech lacks access to a jointly experienced physical context, leading to differences in underlying speech act types.

Although there still remains much work to be done in the space of AI–child interactions, there are already three key contributions made by the present study:

First, the results of our analysis highlight the importance of looking beyond gross metrics of general complexity and diversity when evaluating child-directed output, which may not reflect the learning-relevant properties of language (and thus may underestimate crucial differences between AI- and human-generated speech). Instead, researchers should focus on more fine-grained, linguistically-informed metrics.

Second, our findings raise important considerations for the growing presence of AI toys, educational chatbots, and conversational agents designed for children. Our results caution against replacing caregiver input with AI-generated speech, as there could be potential consequences for child language learning. Further research on child-directed speech and language acquisition, as well as the differences between that of human caregivers and AI, may help elucidate next steps toward mitigating such potential consequences.

Third, our results provide evidence that LLMs are unlikely to serve as effective proxies for human subjects (see also Abdurahman et al., 2024; Crocket & Messeri, 2025), contrary to the arguments made by Argyle et al. (2023) and others. Our work is one of the first studies to investigate this issue with respect to child-directed speech in particular. As such, there is ample future work remaining to further explore the similarities and differences between child-directed speech produced by caregivers and AI (e.g., extending our analysis to investigate the distribution of syntactic structures such as embedded vs. matrix sentences).

In sum, delineating the similarities and differences between caregiver and AI child-directed speech could help us better understand how human-produced input facilitates children's learning, and what steps could improve AI's ability to replicate it, in order to mitigate the potential negative effects of young children's increased exposure to AI.

[4] Note, however, that this contrast persisted even when we changed the goal in the AI system prompt to encourage more discussion of epistemic states: *"Your goal is to share your thoughts and experiences with your child, and engage with theirs."*

Acknowledgments

AI-generated language samples were analyzed as data in this study. Separately, AI-based tools were used for code debugging. All study design, analysis, and interpretation were conducted by the authors. This research was supported by the University of Pennsylvania's Pinkel Undergraduate Research Fund Award and the Data Driven Discovery Initiative Undergraduate Data Science Fellows Grant.

References

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., ... & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS nexus*, *3*(7), 245.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, *31*(3), 337-351.

Axelsson, E. L., Churchley, K., & Horst, J. S. (2012). The right thing at the right time: Why ostensive naming facilitates word learning. *Frontiers in Psychology, 3*, 88.

Bang, J., & Nadig, A. (2015). Learning language in autism: Maternal linguistic input contributes to later vocabulary. *Autism Research, 8*(2), 214–223.

Bickham, D. S., Schwamm, S., Izenman, E. R., Yue, Z., Carter, M., Powell, N., Tiches, K. & Rich, M. (2024). *Use of Voice Assistants & Generative AI by Children and Families.* Boston Children's Hospital Digital Wellness Lab.

Bruner, J. S. (1978). Berlyne Memorial Lecture: Acquiring the uses of language. *Canadian Journal of Psychology / Revue Canadienne de Psychologie, 32*(4), 204–218.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How people use ChatGPT. *National Bureau of Economic Research,* Working paper 34255.

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology, 37*(6), 739.

Crockett, M. J., & Messeri, L. (2025). AI Surrogates and illusions of generalizability in cognitive science. *Trends in Cognitive Sciences*. Advance online publication.

Demuth K., Culbertson J., & Alter J. (2006). Word-minimality, epenthesis and coda licensing in the acquisition of English. *Language and Speech, 49*, 137–174

De Villiers, J. G., & De Villiers, P. A. (1973). A cross-sectional study of the acquisition of grammatical morphemes in child speech. *Journal of Psycholinguistic Research, 2*(3), 267–278.

deWinter, J. C. F. (2013). Using the Student's *t*-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation, 18*(10).

Dillion, D., Tandon, N., Gu, Y., & Gray, K. (2023). Can AI language models replace human participants?. *Trends in Cognitive Sciences*, 27(7), 597-600.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2006). MacArthur-Bates Communicative Development Inventories, Second Edition (CDIs) [Database record]. APA PsycTests.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*(3), 677–694.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology, 45*(3), 337–374.

Kempe, V., Ota, M., & Schaeffler, S. (2024). Does child-directed speech facilitate language development in all domains? A study space analysis of the existing evidence. *Developmental Review, 72*, 101121.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978–990.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Liu, J., & Fourtassi, A. (2024). *Benchmarking LLMs for mimicking child-caregiver language in interaction*. arXiv. https://doi.org/10.48550/arXiv.2412.09318

MacWhinney, B. (2014). *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

Mann, S., Calvin, A., Lenhart, A., & Robb, M. B. (2025). *The Common Sense census: Media use by kids zero to eight, 2025*. Common Sense Media.

Meng, J. (2024). AI emerges as the frontier in behavioral science. *Proceedings of the National Academy of Sciences, 121*(10), Article e2401336121.

OpenAI. (2026). ChatGPT (Jan 28 version) [Large language model]. https://chat.openai.com/chat

Pavliv, V., Akbari, N., & Wagner, I. (2024). AI-powered smart toys: Interactive friends or surveillance devices?. *Proceedings of the 14th International Conference on the Internet of Things*, 172–175.

Portelance, E., Degen, J., & Frank, M. C. (2020). Predicting age of acquisition in early word learning using recurrent neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 42).

Pozzan, L., & Valian, V. (2017). Asking questions in child English: Evidence for early abstract representations. *Language Acquisition, 24*(3), 209–233.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14*(2), 201–209.

Richtsmeier, P., Gerken, L., & Ohala, D. (2011). Contributions of phonetic token variability and word-type frequency to phonological representations. *Journal of Child Language, 38*(5), 951–978.

Roche, E. C., Hirsh-Pasek, K., Romeo, R., Suskind, D., Eulau, K., Zosh, J., Golinkoff, R. M., Evans, N., Yogman, M., Noble, K., Leonard, J., Mackey, A., Carlson, S., Galinsky, E., Duckworth, A., Farah, M., Zelazo, P. D., Shiohira, K., & Garner, A. (2025). Statement on the risks of AI interaction to babies and toddlers around the world.

https://childrensmuseums.org/2025/09/16/scientists-issue-urgent-warning-about-ai-and-young-children/

Snow, C. E. (1977). The development of conversation between mothers and babies. *Journal of Child Language, 4*(1), 1–22.

Wijffels, J., Straka, M., & Straková, J. (2019). Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the 'UDPipe' 'NLP' toolkit

Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. *Big Data & Society, 12*(4), Article 20539517251389860.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language.* MIT press.

Zanotto, S. E., & Aroyehun, S. (2024). *Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models*. arXiv. https://doi.org/10.48550/arXiv.2412.03025

Zhu, A., Dugan, L., Hwang, A., & Callison-Burch, C. (2023). Kani: A lightweight and highly hackable framework for building language model applications. *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, 65–77.