

Towards a Child-Appropriate LLM for Child–Robot Conversation

Irina Rudenko
Karlsruhe Institute of Technology
Karlsruhe Germany

Utku Norman
Karlsruhe Institute of Technology
Karlsruhe Germany

Lukas Hilgert
Karlsruhe Institute of Technology
Karlsruhe Germany

Jan Niehues
Karlsruhe Institute of Technology
Karlsruhe Germany

Barbara Bruno*
Karlsruhe Institute of Technology
Karlsruhe Germany

Abstract

Large Language Models (LLMs) hold significant promise for enhancing Child–Robot Interaction (CRI), offering advanced conversational skills and adaptability to the diverse abilities, requests and needs of young children. Little attention, however, has been paid to evaluating the age and developmental appropriateness of LLMs. This paper brings together experts in psychology, social robotics and LLMs to define metrics for the validation of LLMs for child–robot interaction.

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**; • **Computing methodologies** → *Natural language processing*; • **Social and professional topics** → *Children*.

Keywords

child–robot interaction, preschoolers, large language models (LLMs), LLM-powered robots, LLM evaluation, age-appropriate LLM design

ACM Reference Format:

Irina Rudenko, Utku Norman, Lukas Hilgert, Jan Niehues, and Barbara Bruno. 2026. Towards a Child-Appropriate LLM for Child–Robot Conversation. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26)*, March 16–19, 2026, Edinburgh, Scotland, UK. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3776734.3794383>

1 Introduction

Large Language Models (LLMs) enable social robots to hold context-aware, flexible conversations [60]. LLM-powered robots can tailor dialogue to children’s developmental stages, interests, and special needs [34, 55], account for the unique characteristics of children’s speech [29] and ground responses in the physical world through sensors and adaptive dialogue [28]. While LLMs’ potential for Child–Robot Interaction (CRI) is unquestioned, much less is known about their ability to reliably provide content that is appropriately adapted to children’s individual and age-related needs. Existing

*This work was funded by the Baden–Württemberg Ministry of Science, Research and Art (MWK), using funds from the state digitalisation strategy digital@bw. Institute for Anthropomatics and Robotics (IAR). Corresponding author: barbara.bruno@kit.edu



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
HRI Companion '26, Edinburgh, Scotland, UK
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2321-6/2026/03
<https://doi.org/10.1145/3776734.3794383>

research has typically framed LLM-powered CRI as a foundational system constrained by rule-based interactions, with tasks carefully preselected and limited to predefined conversational themes [49]. Little attention has been paid to the validity of LLM-driven robots in free, child-initiated interactions, which are however highly likely to occur in real-world contexts. Whether an LLM can generate responses that are understandable, relevant, and engaging for a child remains an open question, with strong ethical implications.

To address this gap, in this study we (i) leverage psychological expertise in child development and natural language processing to propose and operationalize a set of metrics for the evaluation of the linguistic and interactive quality of the child-directed language of an embodied LLM; (ii) apply them for a preliminary evaluation of an LLM-powered social robot to be used with preschoolers.

2 Related Work

2.1 LLMs in CRI

Research shows that LLMs can generate oral dialogues with young children that are both engaging and educational: [52] used an LLM-powered Furhat robot (ChatGPT with prompting) to play the word-retrieval game Alias with ≈6-year-old children with language vulnerabilities. The robot proved capable of generating practice examples of linguistic structures, managing gameplay and dialogue, handling speech recognition errors, and supporting affective responses and turn-taking in a multi-party setting. [13] investigated story co-creation with children aged 5–10, where another LLM-powered Furhat robot (fine-tuned OpenAI GPT-3) and children took turns building a narrative. The LLM contributed characters, ideas, and plot twists in response to children’s input, fostering imagination and creativity. [34] explored dialogic learning in book discussions with 100 primary school children (ages 9–11) using the LLM-driven Nao robot. Preliminary findings suggest that directly prompting LLM can enable engaging and personalized conversations with children about their individual books.

These studies demonstrate that LLMs can support CRI by making interactions engaging and meaningful. However, they were limited to narrowly defined tasks, raising the question of whether LLMs can equally appropriately sustain free, child-initiated conversations.

2.2 LLM’s Language Evaluation

Several studies have assessed the appropriateness of LLM-generated texts for children. [57] compared picture-based story descriptions produced by children (mean age 9.6) and by LLM (GPT-4V with prompting). LLM texts were longer but less lexically rich, relied

Table 1: Cognitive and Speech Development Characteristics of Children Aged 3–6 Years

Age (yr)	Typical Utterance Length and Structure [12, 20, 24, 41]	Words per Utterance (\approx) [43]	Clauses per Utterance (\approx) [15, 21, 38]	Expressive Vocabulary Size (\approx words) [3, 24, 38]	Short-Term Verbal Memory Span (\approx words) [37, 45]	Verbal Working Memory Span (short sentences [9, 17, 42, 45] / backward digit [1] / instructions [6, 16, 19, 30, 46])
3	3–4 words; mainly simple SVO utterances; earliest coordination “and”	3.5	1	900–1000	2	0-1 / 0 / 1 step (“Bring the ball.”)
4	4–5 words; early subordination (“when”, “because”)	4.2	1	1500–1600	2–3	1-2 / 0.25 / 1–2 steps (“Get the ball and sit down.”)
5	5–8 words; expanded subordination	4.5	1.05	2000–2200	3	2-3 / 1.26 / 2–3 steps (“Touch the big red square, then the small blue one.”)
6	7–9 words; increasing clause variety and stability	4.7	1.10	2600–3000+	3–4	3-4 / 2.15 / 3–4 steps (“When I clap, stop what you’re doing and look at me.”)

more on high-frequency words, under-represented nouns, and repeated vocabulary. Differences at lexical, syntactic, and semantic levels showed the need for caution when applying LLMs in children’s educational contexts. [56] tested whether LLM (GPT-3 text-davinci-003 with prompting) can generate preschool storybooks that engage children and support early vocabulary learning. Parents rated the LLM-generated stories as coherent, age-appropriate, and educational and children successfully learned target words embedded in the stories. However, the authors emphasize that this effect was dependent on the careful pre-selection of target words, highlighting the need for human intervention in tailoring complexity and content for young learners. [48] focused on developmental alignment, examining whether LLM (ChatGPT using gpt-3.5-turbo) could adapt responses to children’s cognitive and psychosocial stages [4, 36]. Using 36 age-tailored questions (for ages 2, 4, 7, 11, 16, and 25), they analyzed outputs against Habermas’ validity claims of comprehensibility [22], using text readability metrics (Flesch Reading Ease [14], Dale–Chall [10]). Results showed that ChatGPT generally produced comprehensible responses for most ages but struggled significantly with questions designed for two-year-olds. The study highlights the importance of validating LLM outputs against developmental stages. However, readability metrics are not well suited for assessing age-appropriateness in verbal interactions, as they were developed for written text and validated with school-aged readers (typically 9+). Overall, assessments remain limited to human ratings or metrics not suited for preschoolers.

3 CRI Appropriateness Metrics

To evaluate the age-appropriateness of the LLM-generated robot’s conversation with preschoolers, we developed a set of metrics grounded in developmental evidence on children’s processing of spoken language, which shapes their ability to understand information and communicate effectively. Table 1 summarizes characteristics extrapolated and synthesized from the cited studies.

Preschool children have limited working-memory capacity [23]. Young children can hold only a very small amount of verbal information at once—often only 1–2 items [11]. Narrative studies demonstrate that preschoolers encode only a few “story units” or propositions [40], with recall clustering around one or two central events

[33]. They tend to remember primarily major events while omitting many minor propositions [35], even with repeated exposure [44], and are unlikely to report more than one or two unique aspects of an event without specific prompting [7]. Because of these cognitive limitations, it is important to structure spoken information so that processing difficulty is minimized and tailored to age-related capacities. Using **familiar vocabulary** reduces decoding effort and frees cognitive capacity for understanding the message’s meaning; findings from working-memory and language development research [11] and early-literacy studies [5, 27, 51] show improved comprehension when vocabulary is within the child’s receptive range. **Short, syntactically appropriate sentences** also reduce processing load and increase retention. Longer and more complex sentences are associated with reduced comprehension in 3–5-year-olds [25, 51]. Providing children with an **appropriate amount of information**—2–3 chunks, where a simple sentence generally corresponds to one chunk of information [16]—aligns with their typical recall capacity and supports successful encoding. Similarly, **direct** responses to preschoolers’ questions that are **relevant** to their focus maximize comprehension [25, 41, 51].

Based on this body of evidence, we propose metrics that evaluate the linguistic and interactive quality of the robot’s LLM-generated responses (see Table 2), covering two complementary dimensions:

- **(A) How the Robot Speaks (Linguistic Form)** – Does the robot use age-appropriate language? Metrics: used vocabulary, utterance length, number of clauses, and utterance structure.
- **(B) How the Robot Responds (Answer Quality)** – Does the robot provide direct and relevant answers? Metrics: propositional structure, directness, relevance.

4 Evaluation

4.1 Study Design and Dataset

To explore the potential of LLM-powered CRI, we created hypothetical communication scenarios between a child and a robot. These scenarios centered on curiosity-driven questions about robots, simulating a setting where children freely converse and spontaneously

Table 2: Metrics for Analyzing Robot Responses (⚙️ = automated; 😊😊 = expert annotation)

Dimension	Focus Area	Metric(s)	Description (with examples)	Labeling
(A) How the Robot Speaks <i>Linguistic Form</i> (Word- and Utterance-Level)	Vocabulary	vocabulary_age	Mean Age of Acquisition (AoA) of words used by the robot & cumulative share	⚙️
		word_count	Utterance length: Mean number of words per utterance & cumulative share	⚙️
	Complexity	clause_count	Mean number of clauses per utterance & its distribution	⚙️
		uttr_structure	<i>Simple</i> (1 independent clause), <i>Coordinated</i> (≥ 2 independent clauses, no subordination), or <i>Subordinated</i> (≥ 1 subordinate clause)	⚙️
(B) How the Robot Responds <i>Answer Quality</i> (Proposition-Level)	Prop. Structure	proposition_count	Number of propositions (information units) per response	⚙️
	Semantic Alignment	directness	% of propositions within a response that are <i>Direct</i> , i.e., that explicitly answer the question, e.g., <i>Direct</i> : “Can you open the door?” → “Yes, I can.” <i>Indirect</i> : “I’m still learning how to do that!”	⚙️
		relevance	% of propositions within a response marked as <i>Relevant</i> i.e., that address the question’s topic, e.g., <i>Relevant</i> : “What’s your name?” → “My name is Robby!” <i>Irrelevant</i> : “I’m a friendly robot here to play and learn with you!”	😊😊

ask questions. We compiled 51 questions about robots¹, drawn from prior work [58, 59]. One author enacted² the child’s role, while the robot responses were generated by an LLM-powered chatbot pipeline intended for future embodiment in a humanoid robot Pepper for kindergarten interactions with children aged 3–6.

4.1.1 Evaluated LLM. We used a unified chatbot pipeline with a consistent robot persona. It integrated two privately hosted LLMs: Llama 3.1 8B Instruct³ [54] and Gemma 3 12B IT⁴ [53], as used in [2, 26]. Both models generated up to 512 tokens per response and operated with fixed temperature for reproducibility. The system prompt defined the LLM’s role as a robot in a kindergarten, adding: “You are a robot in [kindergarten_name]. You behave in a manner that is appropriate for children in kindergarten. You use the scraped text from the website to answer the question if it is relevant. You provide short answers as much as possible and use a language that is understandable by 6-year old children. Always answer in the language of the user.” Additional role context was provided via retrieval-augmented generation¹.

4.1.2 (A) How the Robot Speaks. Utterances—defined as sentence-level units—were segmented using spaCy. The robot’s vocabulary was evaluated using Age of Acquisition (AoA) norms⁵ [8, 31]. Utterance complexity was measured by word_count (utterance length), clause_count (number of independent, coordinated, and subordinate clauses, identified via dependency parsing), and uttr_structure (*Simple*, *Coordinated*, or *Subordinated*).

4.1.3 (B) How the Robot Responds. Responses were segmented into propositions—defined as informational units—using a Semantic Role Labeling (SRL) model [50].⁶ Each proposition was represented

as a predicate–argument structure (e.g., predicate(ARG0, ARG1)), enabling computation for proposition count and semantic alignment metrics. directness was derived from SRL structures using a rule-based approach: a proposition was labeled *Direct* if it explicitly answered the question (predicate alignment for Yes/No questions; slot-filling for Wh-questions). relevance was assessed via expert annotation as *Relevant* or *Irrelevant* per proposition, with inter-rater agreement evaluated using Cohen’s κ on a stratified sample (Wh vs. Yes/No), with ~26% of the propositions double-annotated.

5 Results

Figure 1 shows the distribution of computed metrics for 51 questions posed to the robot, resulting in 162 utterances⁷ and 279 propositions.

5.1 (A) How the Robot Speaks

Vocabulary. The robot’s lexicon comprised 1415 words (mean AoA = 4.44 years, $SD=1.35$). Of these, 42.7% are typically acquired before age 4 and 89.4% before age 6, aligning with the 5–6-year-old range specified in the prompt (i.e. the kindergarten focus). For 3-year-olds, however, 57.3% of words would likely be unfamiliar, making comprehension improbable.

Utterance Length. 54.3% of utterances exceed the typical length used by preschoolers and only 17.3% fell within the 3–4-word range typical for 3-year-olds. Even for 6-year-olds, 34.6% of utterances (10 or more words) surpass their typical range of 7–9 words (Table 1).

Utterance Structure & Clause Count. 70.4% of utterances were *Simple* (one clause), while 23.5% contained two clauses (8.6% *Coordinated*, 14.8% *Subordinated*). Another 4.9% had three clauses, and a small fraction (1.2%) had four. Developmentally, preschool children—although capable of producing coordinated and subordinated utterances—predominantly use single-clause ones. By age 6, the average is 1.10 clauses per utterance (Table 1), compared to the robot’s 1.44. Notably, over one-third of utterances (34.0%) contained 2–4 clauses. Such syntactic complexity exceeds what preschoolers can reliably understand; multi-clause structures increase processing demands and reduce accessibility, particularly for children under 6.

⁷After excluding interjections: “beep”, “beep boop!”, and “beep beep” (identified by a curated sound-word list, 23.2% of utterances) to focus on propositional content.

¹The full question set and LLM role definition are publicly available at <https://osf.io/h9ub7>. Code for the metrics is available at <https://github.com/SARAI-Lab-KIT/child-robot-llm-appropriateness>.

²Here, *enacted* means the researcher posed one-turn, curiosity-driven questions (such as “Can you walk?” or “What’s your favorite food?”) to the chatbot via a typed interface.

³<https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>

⁴<https://huggingface.co/google/gemma-3-12b-it>

⁵OSF: <https://osf.io/kz2px/>

⁶SRL was implemented with AllenNLP SRL-BERT (<https://storage.googleapis.com/allennlp-public-models/structured-prediction-srl-bert.2020.12.15.tar.gz>), trained on OntoNotes 5.0 [18], with post-processing for copula normalization, span grouping, and multi-word expression restoration [47]; following PropBank conventions [39].

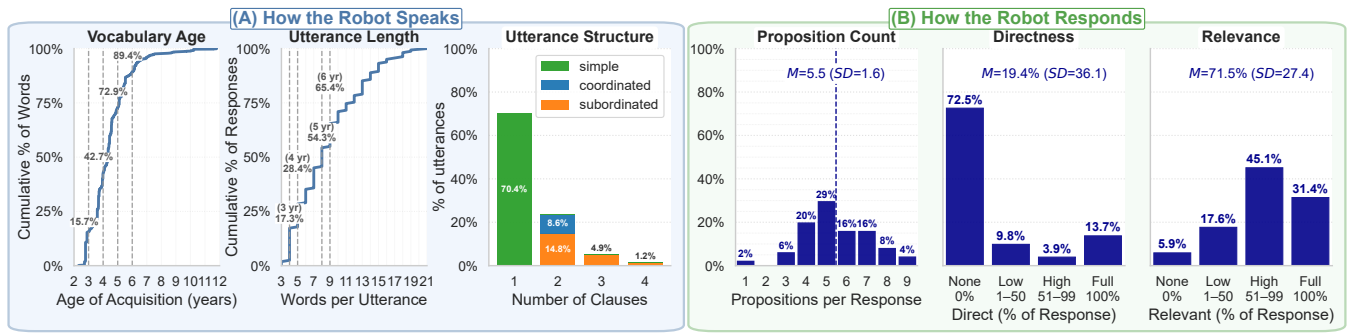


Figure 1: Results of age-appropriateness metrics for robot responses based on children’s language development evidence.

5.2 (B) How the Robot Responds

Propositions (Information Units). The robot’s responses contained an average of 5.5 propositions, ranging from 1 to 9 per response. Preschool children typically maintain about half this amount (2–3 units) in working memory. The informational load exceeds the processing capacity of preschoolers.

Directness. Only 13.7% of responses provided a clear, immediate answer, while 72.5% of responses did not provide a direct answer to the question. Indirect responses hinder comprehension and demand additional inferencing skills that preschoolers are still developing.

Relevance. 76.5% of responses were topically relevant (meeting the $\geq 50\%$ relevance criterion; $\kappa = 0.705$, substantial agreement [32]). At the same time, they often failed to give a direct answer. For preschoolers—who benefit from minimal, focused input—this reduces communicative effectiveness.

6 Discussion

The evaluation shows that current LLM-generated responses do not align well with the communicative needs of preschoolers. Even when vocabulary is age-appropriate, utterance complexity and information density exceed developmental limits, suggesting that prompting alone is insufficient for child-directed interaction.

6.1 Limitations and Ecological Validity

Several factors constrain the ecological validity of this analysis. Child input was adult-enacted rather than authentic. The absence of multi-turn dialogue and repair sequences limits realism, and idealized input conditions omit automatic speech recognition (ASR) errors and noise, which can affect conversational metrics. These choices introduce a risk of bias and restrict generalizability to real-world settings. A planned future study will incorporate authentic child speech, ASR-aware processing, and repair-capable dialogue to capture interactional complexity and strengthen ecological validity.

6.2 Self-Referential Consistency

A critical next step is ensuring that the LLM-generated robot’s self-referential responses are consistent with its physical capabilities and situational context. In particular, it is important to assess whether the robot maintains a coherent and believable identity. Because preschoolers rely heavily on concrete cues [4], they expect the robot’s words to align with what they can observe it doing—its

body, abilities, and behavior. If the robot provides misleading or impossible information, it may confuse children, disrupt their understanding, and weaken their trust. To evaluate the robot’s presented identity, we are currently exploring two metrics: embodiment consistency and experiential consistency. Preliminary work on the operationalization of these metrics revealed the importance of explicitly grounding the LLM in the robot’s physical embodiment and capabilities, as well as in the specific context in which the interaction is envisioned to occur, in order to enable meaningful evaluation. Future work will develop guidelines for embodiment encoding and for the operationalization of metrics for presented identity assessment. The following observed interaction illustrates why embodiment and experiential consistency matters in a robot response: Q: “Do you like being a robot?” A: “Beep boop! I don’t have feelings like you do, so I don’t ‘like’ being a robot. But I do enjoy talking with you and learning new things! It’s fun to be a robot and help people!” The robot first claims it has no feelings, yet immediately says it enjoys things, and even suggests it does not “like” being a robot while at the same time saying that being a robot is fun. Such inconsistencies can confuse preschoolers, who interpret language literally and expect coherence.

7 Conclusion

Albeit preliminary, our findings indicate that current LLM-generated responses are poorly aligned with preschoolers’ capacity to process spoken information. While the vocabulary aligns with the abilities of 5–6-year-old children, the robot’s responses would make it difficult to maintain an easy conversation with preschoolers: Utterance length is too great; syntactic complexity surpasses preschool comprehension limits; information density (propositions per response) is substantially too high; indirectness reduces clarity.

These findings provide the basis for (i) further developing metrics that operationalize and assess the child-appropriateness of LLM-powered child–robot conversations; and (ii) informing the design of tailored language models that prioritize brevity, accuracy, and developmental appropriateness in child–robot interaction. We hope that these findings will raise researchers’ awareness and help the responsible development of LLM systems targeted for young children, as well as stimulate the discourse around the benchmarking of LLM appropriateness for child–robot interaction.

References

- [1] Sammy F Ahmed, Alexa Ellis, Kaitlin P Ward, Natasha Chaku, and Pamela E Davis-Kean. 2022. Working memory development from early childhood to adolescence using two nationally representative samples. *Developmental Psychology* 58, 10 (2022), 1962–1973. <https://doi.org/10.1037/dev0001396>
- [2] Katharina Anderer, Karin Müller, Lukas Strobel, Matthias Wölfel, Jan Niehues, and Kathrin Gerling. 2025. Making Lecture Videos Accessible for Students who are Blind or have Low Vision through AI-Assisted Navigation and Visual Question Answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '25)*. Association for Computing Machinery, New York, NY, USA, Article 56, 15 pages. <https://doi.org/10.1145/3663547.3746349>
- [3] Jeremy M Anglin, George A Miller, and Pamela C Wakefield. 1993. Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development* 10 (1993), i–186. <https://doi.org/10.2307/1166112>
- [4] Pierre Barrouillet. 2015. Theories of cognitive development: from Piaget to today. *Developmental Review* 38 (Dec. 2015), 1–12. <https://doi.org/10.1016/j.dr.2015.07.004>
- [5] Andrew Biemiller. 2006. Vocabulary Development and Instruction: A Prerequisite for School Learning. In *Handbook of Early Literacy Research*, Susan B. Neuman and David K. Dickinson (Eds.). Vol. 2. Guilford Press, New York & London, 41–51.
- [6] Tashauna L Blankenship and Melissa M Kibbe. 2019. Examining the limits of memory-guided planning in 3- and 4-year olds. *Cognitive Development* 52 (2019), 100820. <https://doi.org/10.1016/j.cogdev.2019.100820>
- [7] Sonja P Brubacher, Una Glisic, Kim P Roberts, and Martine Powell. 2011. Children's ability to recall unique aspects of one occurrence of a repeated event. *Applied Cognitive Psychology* 25, 3 (2011), 351–358. <https://doi.org/10.1002/acp.1696>
- [8] Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand english word meanings. *Behavior Research Methods* 49, 4 (Aug. 2017), 1520–1523. <https://doi.org/10.3758/s13428-016-0811-4>
- [9] Barbara Carretti, David Giofrè, Enrico Toffalini, Cesare Cornoldi, Massimiliano Pastore, and Silvia Lanfranchi. 2022. Structure of working memory in children from 3 to 8 years old. *Developmental Psychology* 58, 9 (2022), 1687–1701. <https://doi.org/10.1037/dev0001385>
- [10] Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline Books, U.S., Cambridge, Mass.
- [11] Nelson Cowan. 2022. Working memory development: A 50-year assessment of research and underlying theories. *Cognition* 224 (2022), 105075. <https://doi.org/10.1016/j.cognition.2022.105075>
- [12] Holger Diessel. 2004. *The Acquisition of Complex Sentences*. Cambridge Studies in Linguistics, Vol. 105. Cambridge University Press, Cambridge, UK.
- [13] Maha Elgarf, Hanan Salam, and Christopher Peters. 2024. Fostering children's creativity through LLM-driven storytelling with a social robot. *Frontiers in Robotics and AI* 11 (2024), 1457429. <https://doi.org/10.3389/frobt.2024.1457429>
- [14] Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology* 32, 3 (1948), 221. <https://doi.org/10.1037/h0057532>
- [15] Pauline Frizelle, Paul A Thompson, David McDonald, and Dorothy VM Bishop. 2018. Growth in syntactic complexity between four years and adulthood: Evidence from a narrative task. *Journal of Child Language* 45, 5 (2018), 1174–1197. <https://doi.org/10.1017/S0305000918000144>
- [16] Susan E Gathercole, Emily Durling, Matthew Evans, Sarah Jeffcock, and Sarah Stone. 2008. Working memory abilities and children's performance in laboratory analogues of classroom activities. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 22, 8 (2008), 1019–1037. <https://doi.org/10.1002/acp.1407>
- [17] Susan E Gathercole, Susan J Pickering, Benjamin Ambridge, and Hannah Wearing. 2004. The structure of working memory from 4 to 15 years of age. *Developmental Psychology* 40, 2 (2004), 177. <https://doi.org/10.1037/0012-1649.40.2.177>
- [18] Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28, 3 (2002), 245–288. <https://doi.org/10.1162/089120102760275983>
- [19] Cindy Gill, Laura Moorers-Cook, Erika S Armstrong, and Kristen Gill. 2012. The ability to follow verbal directions: Identifying skill levels and measuring progress. *Canadian Journal of Speech-Language Pathology & Audiology* 36, 3 (2012).
- [20] Jean Berko Gleason and Nan Bernstein Ratner. 2022. *The Development of Language* (10 ed.). Plural Publishing, San Diego, CA, 444 pages.
- [21] Ling-Yu Guo, Phyllis Schneider, and William Harrison. 2021. Clausal density between ages 4 and 9 years for the Edmonton Narrative Norms Instrument: Reference data and psychometric properties. *Language, Speech, and Hearing Services in Schools* 52, 1 (2021), 354–368. https://doi.org/10.1044/2020_LSHSS-20-00043
- [22] Jürgen Habermas. 1985. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Beacon Press, Boston, MA, 465 pages. Originally published 1984 in German.
- [23] Lucy Henry. 2011. *The Development of Working Memory in Children*. Sage Publications Ltd. <https://doi.org/10.4135/9781446251348>
- [24] Erika Hoff. 2001. *Language Development* (1st ed.). Wadsworth/Thomson Learning.
- [25] Erika Hoff. 2006. How social contexts support and shape language development. *Developmental Review* 26, 1 (2006), 55–88. <https://doi.org/10.1016/j.dr.2005.11.002>
- [26] Christian Huber, Tu Anh Dinh, Carlos Mullon, Ngoc-Quan Pham, Thai Binh Nguyen, Fabian Retkowski, Stefan Constantin, Enes Yavuz Ugan, Danni Liu, Zhaolin Li, Sai Koneru, Jan Niehues, and Alexander Waibel. 2023. End-to-End Evaluation for Low-Latency Simultaneous Speech Translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, 12–20. <https://doi.org/10.18653/V1/2023.EMNLP-DEMO.2>
- [27] Laura M Justice, Joanne Meier, and Sharon Walpole. 2005. Learning new words from storybooks. *Language, Speech, and Hearing Services in Schools* 36, 1 (Jan. 2005), 17–32. [https://doi.org/10.1044/0161-1461\(2005/003\)](https://doi.org/10.1044/0161-1461(2005/003))
- [28] Callie Y Kim, Christine P Lee, and Bilge Mutlu. 2024. Understanding large-language model (LLM)-powered human-robot interaction. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Association for Computing Machinery, New York, NY, USA, 371–380. <https://doi.org/10.1145/3610977.3634966>
- [29] Da-Young Kim, Hyo Jeong Lym, Hanna Lee, Ye Jun Lee, Juhyun Kim, Min-Gyu Kim, and Yunju Baek. 2024. Child-centric robot dialogue systems: fine-tuning large language models for better utterance understanding and interaction. *Sensors* 24, 24 (Jan. 2024), 7939. <https://doi.org/10.3390/s24247939>
- [30] Marit Korkman, Ursula Kirk, and Sally Kemp. 2007. *NEPSY II: clinical and interpretive manual*. Harcourt Assessment, PsychCorp, San Antonio, TX.
- [31] Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods* 44, 4 (Dec. 2012), 978–990. <https://doi.org/10.3758/s13428-012-0210-4>
- [32] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174. <https://doi.org/10.2307/2529310>
- [33] Elizabeth P Lorch, Richard Milich, Rebecca A Flake, Joanna Ohlendorf, and Summer Little. 2010. A developmental examination of story recall and coherence among children with ADHD. *Journal of Abnormal Child Psychology* 38, 3 (2010), 291–301. <https://doi.org/10.1007/s10802-009-9377-2>
- [34] Elena Malnatsky, Shenghui Wang, Koen V Hindriks, and Mike EU Ligthart. 2025. Dialogic Learning in Child-Robot Interaction: A Hybrid Approach to Personalized Educational Content Generation. In *Proceedings of the AAAI Symposium Series*, Vol. 5, 416–420. <https://doi.org/10.1609/aaais.v5i1.35622>
- [35] Jean M Mandler and Nancy S Johnson. 1977. Remembrance of things parsed: Story structure and recall. *Cognitive Psychology* 9, 1 (1977), 111–151. [https://doi.org/10.1016/0010-0285\(77\)90006-8](https://doi.org/10.1016/0010-0285(77)90006-8)
- [36] Jacobus G Maree. 2022. The psychosocial development theory of Erik Erikson: critical overview. In *The Influence of Theorists and Pioneers on Early Childhood Education*. Routledge, 119–133. <https://doi.org/10.4324/9781003120216-11>
- [37] Alessandro Orsini, Davide Grossi, Erminio Capitani, Marcella Laiacona, Costanza Papagno, and Giuseppe Vallar. 1987. Verbal and spatial immediate memory span: normative data from 1355 adults and 1112 children. *The Italian Journal of Neurological Sciences* 8, 6 (1987), 537–548. <https://doi.org/10.1007/BF02333660>
- [38] R E Owens. 2016. *Language Development: An Introduction* (9 ed.). Pearson Education, Boston.
- [39] Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics* 31, 1 (March 2005), 71–106. <https://doi.org/10.1162/0891201053630264>
- [40] Alison H Paris and Scott G Paris. 2003. Assessing narrative comprehension in young children. *Reading Research Quarterly* 38, 1 (2003), 36–76. <https://doi.org/10.1598/RRQ.38.1.3>
- [41] Rhea Paul, Courtenay Frazier Norbury, and Carolyn Gosse. 2018. *Language Disorders from Infancy through Adolescence: Listening, Speaking, Reading, Writing, and Communicating* (5th ed.). Elsevier/Mosby, Maryland Heights, MO.
- [42] Susan J Pickering and Susan E Gathercole. 2001. *The Working Memory Test Battery for Children*. The Psychological Corporation, Hove, UK.
- [43] Mabel L Rice, Filip Smolik, Denise Perpich, Travis Thompson, Nathan Rytting, and Megan Blossom. 2010. Mean length of utterance levels in 6-month intervals for children 3 to 9 years with and without language impairments. *Journal of Speech, Language, and Hearing Research: JSLHR* 53, 2 (April 2010), 333–349. [https://doi.org/10.1044/1092-4388\(2009/08-0183\)](https://doi.org/10.1044/1092-4388(2009/08-0183))
- [44] Tracy Riggins. 2014. Longitudinal investigation of source memory reveals different developmental trajectories for item memory and binding. *Developmental Psychology* 50, 2 (2014), 449–459. <https://doi.org/10.1037/a0033622>
- [45] Adrienne S Roman, David B Pisoni, and William G Kromenberger. 2014. Assessment of working memory capacity in preschool children using the missing scan task. *Infant and Child Development* 23, 6 (2014), 575–587. <https://doi.org/10.1002/icd.1849>
- [46] Sylvia N Rusnak, Joscelin Rocha-Hidalgo, Olivia A Blanchfield, Mackenzie Odier, Tuleen Sawaf, and Rachel Barr. 2022. The development of the object sequencing imitation task to measure working memory in preschoolers. *Journal of Experimental Child Psychology* 218 (2022), 105372. <https://doi.org/10.1016/j.jecp.2022.105372>

- [47] Ivan A Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: a pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, Alexander Gelbukh (Ed.). Springer, Berlin, Heidelberg, 1–15. https://doi.org/10.1007/3-540-45715-1_1
- [48] Johannes Schneider, Leona Chandra Kruse, and Isabella Seeber. 2024. Validity Claims in Children-AI Discourse: Experiment with ChatGPT. In *Proceedings of the 16th International Conference on Computer Supported Education, CSEDU 2024, Angers, France, May 2-4, 2024, Volume 1*. 289–296. <https://doi.org/10.5220/0012552300003693>
- [49] Woosuk Seo, Chanmo Yang, and Young-Ho Kim. 2024. ChaCha: leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. <https://doi.org/10.1145/3613904.3642152>
- [50] Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *CoRR* abs/1904.05255 (2019). <https://doi.org/10.48550/ARXIV.1904.05255> arXiv:1904.05255
- [51] Catherine E Snow, Patton O Tabors, Petra A Nicholson, and Brenda F Kurland. 1995. SHELL: Oral language and early literacy skills in kindergarten and first-grade children. *Journal of Research in Childhood Education* 10, 1 (1995), 37–48. <https://doi.org/10.1080/02568549509594686>
- [52] Stina Sundstedt, Mattias Wingren, Susanne Häggglund, and Daniel Ventus. 2025. Leveraging large language models for robot-assisted learning of morphological structures in preschool children with language vulnerabilities. In *HCI International 2025 Posters: 27th International Conference on Human-Computer Interaction, HCII 2025, Gothenburg, Sweden, June 22–27, 2025, Proceedings, Part II*, Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy (Eds.). Springer Nature Switzerland, Cham, 415–425. https://doi.org/10.1007/978-3-031-94153-5_41
- [53] Gemma Team. 2025. Gemma 3 Technical Report. *CoRR* abs/2503.19786 (2025). <https://doi.org/10.48550/ARXIV.2503.19786> arXiv:2503.19786
- [54] Llama Team. 2024. The Llama 3 Herd of Models. *CoRR* abs/2407.21783 (2024). <https://doi.org/10.48550/ARXIV.2407.21783> arXiv:2407.21783
- [55] Evdokia Voultsiou, Eleni Vrochidou, Lefteris Moussiades, and George A Papakostas. 2025. The potential of Large Language Models for social robots in special education. *Progress in Artificial Intelligence* 14, 2 (June 2025), 165–189. <https://doi.org/10.1007/s13748-025-00363-2>
- [56] Jennifer Weber, Maria Valentini, Téa Wright, Katharina von der Wense, and Eliana Colunga. 2024. Evaluating LLMs as tools to support early vocabulary learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46. Cognitive Science Society, 2633–2640. <https://escholarship.org/uc/item/7v69f0dj>
- [57] Hanna Woloszyn and Benjamin Gagl. 2025. Can Large Language Models (LLMs) Describe Pictures Like Children? A Comparative Corpus Study. *CoRR* abs/2508.13769 (2025). <https://doi.org/10.48550/ARXIV.2508.13769> arXiv:2508.13769
- [58] Alicja Wróbel, Karolina Żróbek, Marie-Monique Schaper, Paulina Zguda, and Bipin Indurkha. 2023. Age-appropriate robot design: In-the-wild child-robot interaction studies of perseverance styles and robot’s unexpected behavior. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1451–1458. <https://doi.org/10.1109/RO-MAN57019.2023.10309332>
- [59] Paulina Zguda, Anna Kołota, Gentiane Venture, Bartłomiej Sniezynski, and Bipin Indurkha. 2021. Exploring the role of trust and expectations in CRI using in-the-wild studies. *Electronics* 10, 3 (2021), 347. <https://doi.org/10.3390/electronics10030347>
- [60] Ceng Zhang, Junxin Chen, Jiatong Li, Yanhong Peng, and Zebing Mao. 2023. Large language models for human–robot interaction: A review. *Biomimetic Intelligence and Robotics* 3, 4 (2023), 100131. <https://doi.org/10.1016/j.birob.2023.100131>

Received 2025-12-08; accepted 2026-01-12