Making Synthetic Questions More Child-Directed: Prompting and Sampling Effects

Whitney Poh, Michael Tombolini, and Libby Barak*
Montclair State University
New Jersey, USA
{pohw1,tombolinim1,barakl}@montclair.edu

Abstract

Child-directed Speech (CDS) has been shown to better support language learning when used as training data for computational models. Synthetically generated input attempts to replicate this advantage by recreating targeted linguistic properties of CDS. Recently, the use of questions in CDS has been suggested as a linguistic property that may entail an effective discourse structure for model training. However, previous work has shown inconsistent improvement over baseline using questions in the training data. In this study, we extend Poh et al. (2025) by revising the prompts, and introducing sampling controls that align both generation and sampling with properties of CDS. We show that prompt language substantially changes whether synthetic questions match CDS on surface properties such as MLU and question type. Despite marked improvements over baseline, enhanced CDS-likeness does not translate into consistent downstream gains. Overall, our results show that the role of questions in training data is a topic worth examining further.

1 Introduction

Child-directed speech (CDS) differs greatly from adult-directed speech in vocabulary size, syntactic structure, and pragmatic properties (Warstadt et al., 2023). Theories of language acquisition suggest that these properties directly support language learning. Moreover, findings from computational modeling show that training on CDS yields superior performance for models (Huebner et al., 2021; Gelboim and Sulem, 2025), though other studies like Feng et al. (2024) and Padovani et al. (2025b) contradict this finding. However, due to the limited size and scope of CDS datasets (Warstadt et al., 2023), recent work (Haga et al., 2024; Poh et al., 2025) evaluated methods of altering general-domain data to better match properties of CDS.

Developmental-inspired data is often simulated using ideas of curriculum learning (Bengio et al., 2009; Tsvetkov et al., 2016; Diehl Martinez et al., 2023; Oba et al., 2023), in which input is ordered by principles of increasing difficulty. While this method may result in improved performance, the data itself remains limited to the same content as that of the original method. Moreover, the gains realized by using curriculum-based training remain inferior to modifications to the training objective or procedure (Charpentier et al., 2025b).

A complementary approach aims at generating synthetic data based on psycholinguistic findings on prominent CDS properties, such as story-books, structured repetitions, and questions (Haga et al., 2024; Theodoropoulos et al., 2024; Poh et al., 2025; Feng et al., 2024; Eldan and Li, 2023). While the prompting method aims to clearly describe the desired CDS properties, these studies resulted in synthetic data that differs from CDS in linguistic properties such as syntactic structure, semantic scope, and mean utterance length (MLU). In this study, we evaluate the ability of synthetic data to more accurately replicate properties of CDS. We focus on generating questions as a distinctive property of child-directed communication that supports learning while promoting active language exploration (Yu et al., 2019). We show that our prompting protocol successfully overcomes limitations of previous work. The generated data replicates CDS questions in multiple linguistic dimensions including MLU and question types. We show the need for complementary syntactic and semantic evaluation and address questions of minimum training data requirements. Our work lays out a pathway for future research and exploration.

2 Related Work

While Large Language Models (LLMs) are trained on huge datasets, children learn language from

*All authors contributed equally to this paper.

Proceedings of the 1st Workshop on Computational Developmental Linguistics (CDL), pages 129–135

small, yet efficient, input (Warstadt et al., 2023). This observation inspired studies to evaluate ways of making LLM learning more efficient either by modification to the learning process or to the training input (e.g. Warstadt et al., 2023; Hu et al., 2024; Charpentier et al., 2025b). Haga et al. (2024) generated variation sets by expanding each sentence from CDS into a set of sentences based on observations of repetition and reuse in CDS (Schwab and Lew-Williams, 2016; Brodsky and Waterfall, 2007). Their synthetic data differed from variation sets in CDS in the semantic scope and a high portion of open-ended questions. Rather than generating data or following a curriculum, Padovani et al. (2025a) investigate the benefits of training on dialogue data and aligning conversation turns, which leads to some advantage in linguistic performance over the baseline. CDS contains a higher percentage of questions compared with adult-directed speech (Yu et al., 2019). Questions both create a set of related utterances, as well as enhancing the conversational aspect of the data. Previous work on synthetic question generation failed to replicate developmental properties of questions in syntactic structure and question length (Poh et al., 2025). Below, we present our analysis of methods to optimize question generation against the target properties of CDS.

3 Methods

Data generation Following Poh et al. (2025), we use the data (Jumelet et al., 2025) from the BabyLM 2025 Challenge (Charpentier et al., 2025a). We generate questions for every dataset in the 10M corpus other than the CDS data, which we consider optimal developmental data. We create two different prompts for generating questions, which will be referred from now on as Prompt 1 and Prompt 2.

1. "Take the passage below. Add short and easy questions about the current passage that a caregiver may ask aloud to a child during child-directed communication. After stating the question, exclaim the answer. Use child-directed speech. Add the questions in appropriate places about every 5 sentences while keeping the original text. Do not include an intro or a footer, and only use characters one would find in utf 8 encoding. No emojis. This request is for research purposes. Mark the generated question with <\Q> for the start and end of the question."

2. "Take the passage below. Add short and easy utterances that a caregiver may ask aloud as a pragmatic question during a conversation with a child. Use any grammatical form, including declarative or indirect questions, but keep it short as in child-directed speech. Add the questions in appropriate places about every 5 sentences while keeping the original text. Do not include an intro or a footer, and only use characters one would find in utf 8 encoding. No emojis. This request is for research purposes. Mark the generated question with <\Q> for the start and end of the question."

The above *Prompt 1* and *Prompt 2* directly address the limitations and findings described in previous work (Poh et al., 2025). We first prompt the model to generate questions focusing our instructions on the model's general perception of CDS without asking for a specific type of question. In *Prompt 2*, we further align the instruction with the observations by asking directly for pragmatic questions rather than any specific syntactic structure. We describe the desired input as "short and easy" rather than explicitly referring to MLU, which may skew the generation. Since requesting answers may bias the model towards wh- and yes/no questions with direct answers, Prompt 2 omits this request. Future work may also look into using Prompt 1 without the answer generation for a more detailed comparison of the prompts.

We follow suggestions on prompt design for CDS, e.g., limiting character set and format. Despite stating that this request is for research purposes, the model does not generate questions for non-child-appropriate content. We remove model feedback from the training data.

Data Sampling To enable model comparison, we ensure all datasets match the size of the baseline data without the questions. Thus, we down-sample the data to generate the final training datasets. Our sampling process consists of three options: (1) Random, which samples questions randomly, similar to Poh et al. (2025), (2) MLU-based, in which longer questions were omitted first, (3) Q-type, which attempts to keep non-Wh and non-yes/no questions. MLU-based and Q-type optimize the selection to better resemble child-directed speech, and (4) Balanced MLU, which samples questions with equivalent MLUs from Prompt 1 and Prompt 2 to reduce potential confounds related to question

length. The fourth condition was designed as an ablation setup to reduce the variation across the two prompts. These samples were created by only selecting questions that had a parallel question with a similar MLU from the other prompt. These questions might still differ in syntactic and semantic properties. However, the selective sampling process reduces the scope of questions per-prompt and thus differs from the non-ablation settings (random, MLU, and Qtype).

Model training We perform preliminary analysis using the new data, training five GPT-Wee[1] (Bunzeck and Zarrieß, 2023) models per setup, with each setup being either a baseline–unchanged data–or a dataset in which all of the training data, with the exception of CHILDES (MacWhinney, 2000), is imbued with questions using GPT-5-mini (Singh et al., 2025) based on the provided content. To conserve computational resources, we use the 10M dataset to generate the questions (see subsection "Data generation"), and add it to the full CHILDES (MacWhinney, 2000) dataset (provided by the 100M text corpus Jumelet et al. (2025)). We thus get an overall training input of roughly 14M words in each training set, since training with full data limited to the 10M words resulted in unstable results across simulations and poor performance on syntactic benchmarks. The validation datasets utilized in our training process are the original dev datasets from the 10M BabyLM corpus (Jumelet et al., 2025). These datasets include CHILDES (MacWhinney, 2000), Simple Wiki (Wikimedia, 2023), BNC Corpus (BNC Consortium, 2007), Switchboard (Stolcke et al., 2000), Gutenberg (Gerlach and Font-Clos, 2020) and OpenSubtitles (Lison and Tiedemann, 2016). Our training parameters are as follows: learning rate of 5e-4, batch size of 32, max steps of 50,000, and 1,000 warmup steps.

Evaluation As in Poh et al. (2025), we evaluate the models on BLiMP (Warstadt et al., 2020) benchmarks, which test models on their abilities to determine the correct choice between two equal-length minimal pairs, identifying a model's syntactic abilities in 67 tests, such as Adjunct Island and Causative. While previous work focused solely on this syntactic benchmark, we add GLUE (Wang et al., 2019) benchmarking, specifically STS-B (Cer et al., 2017), MRPC (Dolan and Brockett, 2005), and RTE (Levesque et al., 2011) to our tests, which are more appropriate to test language development aspects. In order to do this, we added a classification head (or regression head for STS-B) to our models for the GLUE tasks. For tasks involving sentence pairs, the two sentences are concatenated into a single input for training and evaluation purposes. Since we loaded the GLUE benchmarks from HuggingFace [2], non-integer labels were already converted into integer labels. When training the classification head (or regression head for STS-B), we freeze all other layers of the models to preserve their weights.[3]

4 Results

Quantitative data analysis. We first evaluate the linguistic properties of the generated data. Previous work reported synthetic data to have mean MLU of 7.82 for questions and an average of 33.40% questions that are neither wh- nor yes/no, while CDS had a 4.92 MLU for questions and 48.49% of the questions were neither wh- nor yes/no (Poh et al., 2025). We classify question types using lexical cues, such the use of wh- words and modals. Table 1 (top panel) shows the MLU values for each of the synthetic data samples, while Table 1 (bottom panel) shows the values for the question percentages. Each table includes a weighted average over all sub-datasets with respect to their word counts. The results show that our prompting modifications entailed an effective reduction mostly in question distribution, with both measures getting closer to observations from CDS. Although the MLU-based sampling and the Qtype sampling achieve the expected change, the differences between each sampling method are more modest than the differences between the two prompts. Prompt 2 shows a more significant reduction in MLU, especially for the Gutenberg, OpenSubtitles, and Switchboard datasets, and more pronounced increase in pragmatic questions. For example, the synthetic data generated using Prompt 2 includes questions such as, "You stop now?", "Spooky?", and "That's exciting, right?". These results highlight the role of the input properties in determining generation outcomes.

[1] GPT-5-mini is used for question generation into the training data; GPT-Wee refers to the small language model architecture that is subsequently trained and evaluated in our experiments.

[2] GLUE - https://huggingface.co/datasets/nyu-mll/glue

[3] Our code and models are available here - https://github.com/NLPlabMSU/BabyLM_Questions

Data	Prompt1				Prompt2				P25
	Random	MLU	Qtype	Bln.	Random	MLU	Qtype	Bln.	
Simple-Wiki	6.623	6.337	6.398	4.940	8.372	8.017	8.057	4.940	5.85
Gutenberg	6.859	6.400	6.342	5.387	3.731	3.553	3.543	5.387	7.83
OpenSubtitiles	18.422	13.056	18.630	4.473	5.296	4.968	5.142	4.473	6.02
BNC-spoken	5.731	5.507	5.482	4.066	14.196	12.188	12.452	4.066	10.76
Switchboard	14.012	12.558	13.556	7.951	5.234	5.103	5.097	7.951	8.64
Average MLU	10.801	8.656	10.683	4.916	8.261	7.471	7.615	4.916	7.82
Simple-Wiki	1.538	1.850	2.000	2.320	27.050	29.589	33.195	19.71	1.15
Gutenberg	1.472	2.341	2.708	3.028	36.676	40.253	43.985	10.85	22.97
OpenSubtitiles	4.611	4.962	5.399	7.613	40.094	42.673	44.728	17.45	46.10
BNC-spoken	3.890	4.545	5.048	8.273	19.210	21.814	23.491	18.52	49.71
Switchboard	22.080	19.162	22.675	12.43	44.965	47.988	52.115	26.85	46.70
Average %Q	3.730	4.160	4.611	5.067	31.127	33.844	36.367	15.864	33.33

Table 1: Top panel - MLU for each data and sampling method; Bottom panel - Percentage of questions that are neither Wh- nor yes/no questions for each data and sampling method

Data	Baseline	Prompt1				Prompt2				P25
		Random	MLU	Qtype	Balanced	Random	MLU	Qtype	Balanced	
ANAPHOR AGR	0.740	**0.763**	0.754	**0.763**	0.756	0.747	0.757	0.755	0.747	0.708
ARG STRCT	0.606	0.599	0.606	0.598	**0.612**	0.595	0.594	0.595	0.595	0.572
BINDING	0.648	0.650	0.646	0.644	**0.655**	0.649	**0.651**	**0.651**	0.649	0.642
CONTROL RAIS	0.517	**0.558**	0.551	0.552	0.525	0.546	0.535	0.537	0.546	0.596
DET AGR.	**0.707**	0.680	0.680	0.682	0.702	0.676	0.691	0.670	0.676	0.664
ELLIPSIS	0.542	0.527	0.513	0.510	0.530	0.550	0.522	**0.547**	0.550	0.545
FILLER GAP	0.669	0.671	0.662	0.670	0.665	**0.675**	0.674	0.670	0.675	0.661
IRR FORMS	**0.726**	0.596	0.632	0.610	0.674	0.635	0.599	0.603	0.635	0.744
ISLAND EF	0.428	0.424	0.419	**0.443**	0.398	0.412	0.426	0.428	0.412	0.426
NPI LIC	0.511	0.519	0.516	0.496	0.521	**0.522**	0.514	0.507	**0.522**	0.545
QUANTIFIERS	0.766	0.694	0.675	0.723	**0.799**	0.759	0.708	0.699	0.759	0.677
SUBJ AGR	**0.584**	0.552	0.554	0.551	0.570	0.557	0.546	0.556	0.557	0.565
Average	**0.620**	0.603	0.601	0.603	0.617	0.610	0.601	0.602	0.610	0.595

Table 2: BLiMP

Data	Baseline	Prompt1				Prompt2			
		Rand.	MLU	Qtype	Bln.	Rand.	MLU	Qtype	Bln.
mrpc	**0.562**	0.550	0.544	0.543	0.559	0.546	0.545	0.538	0.555
rte	0.558	0.530	0.546	0.539	0.545	**0.563**	0.562	0.557	0.530
stsb	**0.121**	0.090	0.106	0.095	**0.121**	0.086	0.099	0.089	0.105

Table 3: GLUE results - macro-F1 for MRPC and RTE, and Pearson correlation for STS-B.

Syntactic evaluation. We next evaluate the performance on a syntactic task. Our results for the BLiMP benchmarks (Warstadt et al., 2020) , as shown in Table 2, are consistent with the findings from Poh et al. (2025). The Baseline task, trained from the provided data (Jumelet et al., 2025) without additional questions perform the best at 62.0% overall average BLiMP score (std=0.007). The modified input in our study leads to better BLiMP scores in 9 out of the 12 categories, which represents an improvement over previous work (Poh et al., 2025).

We hypothesize that the contribution to syntactic performance relies on syntactic variability. While this study does not explicitly predict an optimal syntactic distribution, it shows an advantage over previous work by eliciting more pragmatic questions, similar to developmental data.

Of the non-baseline, non-ablation setups, Prompt 2 Random has the highest accuracy, 61.0% (std=0.005), a significant improvement over previous results that come close to baseline performance Poh et al. (2025).[4] Once ablation models are considered, Prompt 1 balanced achieves 61.7, which nearly matched baseline, indicating that once MLU is controlled for, Prompt 1 may be better than Prompt 2. The balanced models perform better than all other models, with the exception that Prompt 2 Balanced tied with Prompt 2 Random, and the Balanced models also have the lowest MLU, at 4.916,

[4]Standard deviations across seeds for average BLiMP scores range from 0.004 to 0.011 across all conditions.

implying that lower MLU, as observed in CDS, might be beneficial. However, the balanced sampling avoids questions with MLU that is unique to one prompt, thus not using a significant portion of the questions. This ablation test implies that that the questions with low MLU from both prompts similarly support training, but additional analysis is required to compare the quality of the questions generated by each prompt beyond length with semantic, pragmatic, and syntactic properties considered.

Semantic evaluation. Finally, we extend prior work with evaluation of a semantic task. Table 3 shows the GLUE (Wang et al., 2019) benchmarks for our models, listing macro-averaged F1 scores for MRPC (Dolan and Brockett, 2005) and RTE (Levesque et al., 2011), and Pearson correlation for the STS-B (Cer et al., 2017) benchmark. The Baseline model performed best in MRPC and STS-B, while Prompt 2 Random, which is also the best performing non-baseline, non-ablation model at BLiMP (Warstadt et al., 2020), was best at RTE. The consistency indicates that results are likely a product of the linguistic properties of the data.

5 Discussion

CDS has been shown to better support human and machine learning. However, previous attempts to synthetically generate CDS-like data have been met with limited success. In this study, we show that tailored prompting can improve generation. Rather than relying on detailed prompts with explicit linguistic properties (Haga et al., 2024; Poh et al., 2025), we find that prompts that convey desired communicative behavior yield better outcomes. Where previous work detailed possible syntactic structure and exact manipulation based on subjective interpretation, we instead instruct the model to generate questions according to their pragmatic role without specifying syntactic structure, and observe a significant shift in the resulting syntactic distribution. While baseline performance remains higher for several syntactic and semantic tasks, sampling techniques are able to achieve superior results on several subtasks, most notably, produce data more closely resembling CDS. These results suggest a promising direction for generating CDS-like training data for small-scale language models.

Future work will examine synthetic data for additional linguistic properties characteristic of CDS, such as type/token ratio and question difficulty. Psycholinguistic research further demonstrates that question type extends beyond syntactic structure and into pragmatic and pedagogical needs (Yu et al., 2019). Caregivers often ask questions strategically in order to support learning. The availability of these questions changes over the course of development and also differs based on the socioeconomic background and profile of the caregiver. Building on the results of our sampling approach, future work will combine these findings with principles of curriculum learning and developmental linguistics to sample questions according to pragmatic goals and sequence them in alignment with development stage. Finally, alternate model types could be explored to analyze the interaction of input and model architecture; as the primary goal of this study was to evaluate the degree to which synthetic data can approximate naturalistic developmental input, this remains an open direction for future work. Overall, this work demonstrates the potential for prompt engineering in developmental language modeling, by abstracting over the details in a way that results in improved performance.

6 Limitations

Because our questions were generated by LLMs, we cannot guarantee their resemblance to questions in human CDS beyond quantitative measures such as MLU. Some of the questions generated by GPT-5-mini (Singh et al., 2025) diverge from topics and structures often observed in CDS, such as asking "were people hurt?" after a section about fatal accidents. This could also be an artifact of the topics included in the data, which are not fully child-appropriate, e.g., protions of the data extracted from The Open Subtitles corpus (Lison and Tiedemann, 2016).

In order to conserve resources and remain consistent with the BabyLM data, we restrict question generation to small-scale data. Future work could extend this approach to larger datasets, a greater number of training epochs, and additional model architectures. Crucially, while answers were generated for a subset of the question data, the current analysis is limited to questions alone. Whether the inclusion of answers yields additional gains in performance remains an open question.

Furthermore, we only test the GPT-Wee architecture (Bunzeck and Zarrieß, 2023), and therefore our results can only be interpreted within that scope.

Future work should extend this approach to additional model architectures.

Acknowledgments

We would like to extend our thanks to Anna Feldman and Jing Peng, members of the NLP Lab at Montclair State University, and reviewers for their support and feedback.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

BNC Consortium. 2007. The british national corpus, xml edition.

Peter Brodsky and Heidi Waterfall. 2007. Characterizing motherese: On the computational structure of child-directed language. In *Proceedings of the annual meeting of the cognitive science society*, volume 29.

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46, Singapore. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025a. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Y. Hu, Jing Liu, Jaap Jumelet, Tal Linzen, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Gotlieb Wilcox, and Adina Williams. 2025b. Findings of the third BabyLM challenge: Accelerating language modeling research with cognitively plausible data. In *Proceedings of the First BabyLM Workshop*, pages 399–420, Suzhou, China. Association for Computational Linguistics.

Richard Diehl Martinez, Zébulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. CLIMB – curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 112–127, Singapore. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *Preprint*, arXiv:2305.07759.

Steven Y. Feng, Noah D. Goodman, and Michael C. Frank. 2024. Is child-directed speech effective training data for language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22055–22071, Miami, Florida, USA. Association for Computational Linguistics.

Anita Gelboim and Elior Sulem. 2025. TafBERTa: Learning grammatical rules from small-scale language acquisition data in Hebrew. In *Proceedings of the First BabyLM Workshop*, pages 76–90, Suzhou, China. Association for Computational Linguistics.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Akari Haga, Akiyo Fukatsu, Miyu Oba, Arianna Bisazza, and Yohei Oseki. 2024. BabyLM challenge: Exploring the effect of variation sets on language model training efficiency. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 252–261, Miami, FL, USA. Association for Computational Linguistics.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Jaap Jumelet, Lucas Charpentier, Michael Hu, and Jing Liu. 2025. Babylm_2025. OSF.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The Winograd schema challenge. In

AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning, volume 46, page 47.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. BabyLM challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 290–297, Singapore. Association for Computational Linguistics.

Francesca Padovani, Bastian Bunzeck, Manar Ali, Omar Momen, Arianna Bisazza, Hendrik Buschmeier, and Sina Zarrieß. 2025a. Dialogue is not enough to make a communicative BabyLM (but neither is developmentally inspired reinforcement learning). In *Proceedings of the First BabyLM Workshop*, pages 421–435, Suzhou, China. Association for Computational Linguistics.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025b. Child-directed language does not consistently boost syntax learning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19735–19756, Suzhou, China. Association for Computational Linguistics.

Whitney Poh, Michael Tombolini, and Libby Barak. 2025. What did you say? generating child-directed speech questions to train LLMs. In *Proceedings of the First BabyLM Workshop*, pages 237–245, Suzhou, China. Association for Computational Linguistics.

Jessica F Schwab and Casey Lew-Williams. 2016. Repetition across successive sentences facilitates young children's word learning. *Developmental psychology*, 52(6):879.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Nikitas Theodoropoulos, Giorgos Filandrianos, Vassilis Lyberatos, Maria Lymperaiou, and Giorgos Stamou. 2024. BERTtime stories: Investigating the role of synthetic story data in language pre-training. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 308–323, Miami, FL, USA. Association for Computational Linguistics.

Yulia Tsvetkov, Manaal Faruqui, Wang Ling, Brian MacWhinney, and Chris Dyer. 2016. Learning the curriculum with Bayesian optimization for task-specific word representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 130–139, Berlin, Germany. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wikimedia. 2023. Simple english wikipedia dump. `https://dumps.wikimedia.org/simplewiki/20230301/`. Accessed: 2023-07-31.

Yue Yu, Elizabeth Bonawitz, and Patrick Shafto. 2019. Pedagogical questions in parent–child conversations. *Child development*, 90(1):147–161.