What did you say? Generating Child-Directed Speech Questions to Train LLMs

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

Abstract

Child-Directed Speech (CDS) holds unique linguistic properties that distinguish it from other types of textual corpora. Language models trained using CDS often obtain superior results compared with the same size of different types of data. Several studies have aimed at modifying non-CDS data to mimic its linguistic properties to match the hypothesized advantageous aspects of CDS. Here, we propose to adapt the non-CDS portions of the training data to include questions similar to CDS interaction. We modify the data by adding artificially generated questions to the data and methodically analyzing the change in performance using each modified dataset. Our results show that artificial question generation strongly depends on the properties of the original dataset. While the performance improves for question-related measures, the overall performance is negatively affected as a result of the reduced syntactic diversity.

1 Introduction

Child-Directed Speech (CDS) records dialogues between adults and children over daily activities, free play, book readings, etc. Like other conversational text data, CDS follows turn-taking social interaction within a shared context. At the same time, CDS differs from adult-to-adult speech in various linguistic aspects, such as shorter sentences, limited types of grammatical constructions, and a limited number of word types (Cameron-Faulkner et al., 2003). Despite this seemingly reduced complexity, language models have achieved better performance using CDS as training data compared with the same-sized data from other domains (You et al., 2021; Mueller and Linzen, 2023a). Following such findings, previous studies have aimed to mimic the linguistic properties of CDS to evaluate their contribution to the model performance (Tsvetkov et al., 2016; Edman and Bylinina, 2023; Haga et al., 2024). Here, we focus on one such aspect of CDS, namely the high frequency of questions, and analyze how increasing the rate of CDS-like questions affects model performance.

Compared with adult-directed conversations, psycholinguistic studies have found increased frequency of questions in CDS (Cameron-Faulkner et al., 2003; Newport et al., 2020). Such questions follow a formulaic structure that may be beneficial for language acquisition, often starting with the same word sequence, e.g., *"What did..."* and *"Are you..."* (Cameron-Faulkner et al., 2003). From a pragmatic point of view, questions serve various communication goals, such as an opportunity for clarification, verification, and as an attention getter (Rowe, 2008; Callanan and Oakes, 1992). Given the turn-taking nature of the conversation, questions expand on a current topic, creating a semantic flow, possible word overlap, and a diverse set of constructions all relating to a shared topic. These sets of sentences may create repetitions across successive sentences, i.e., variation sets, which have been shown to support the language acquisition of children and possibly the training of computational models (Schwab and Lew-Williams, 2016; Brodsky and Waterfall, 2007; Haga et al., 2024). While questions serve social and pedagogical goals in natural communication, the linguistic properties of this conversational tool may explain how it can support language model training from a computational perspective.

In this study, we look into the role of child-appropriate questions by extending the datasets included in the *Strict-Small* data with artificial child-directed questions (Hu et al., 2024). We first analyze the use of questions in all subsets of the provided datasets. We use GPT-5 (OpenAI, 2025) to generate artificial child-directed questions for each of the data sources. Since the generation of questions increases the overall size of the data, we

*All Authors contributed equally to this paper.

Proceedings of the First BabyLM Workshop, pages 237–245

down-sample each data set independently to maintain the same size of training data as the original data, while preserving the communicative sequence of the questions. We evaluate the contribution of question asking per each data source by methodically constructing versions of the training data that replace one data set at a time with the same data source with artificial questions.

Our results show that, contrary to expectations, most of the data sources provided as part of the original training data include a significant number of questions. However, we find that the linguistic properties of the questions differ from those observed in the data taken from CHILDES (MacWhinney, 2000). Moreover, we find that the question generation varies significantly across datasets depending on the linguistic properties of the original dataset. Finally, the data enhanced with the question data results in better learning of tasks related to the grammatical constructions of questions. However, the overall performance for other linguistic categories decreases. We provide qualitative and quantitative analysis that illustrates how artificial data generation can be a double-edged sword when the generated linguistic properties diverge from natural language and offer directions for future research.

2 Related Work

2.1 Questions in Child-Directed Speech

The use of questions encourages children to become active in their learning, to engage in turn-taking, and produce more language (Snow and Ferguson, 1977). While Yes/No questions can be used as an attention getter and verification of understanding, Wh-questions expose children to more complex syntactic structures. Cameron-Faulkner et al. (2003) find that children repeat the same structures observed in CDS in the language produced by children. They conclude that the repeated expression through the formulaic question pattern supports the learning of complex grammar and models its use in language.

Previous papers have discussed differences between two types of questions–information-seeking questions and pedagogical questions (e.g. Bascandziev et al., 2021), which are questions to which the asker knows the answer, asked for the purpose of teaching or bringing attention to an intended target (Daubert et al., 2020; Jean et al., 2019; Yu et al., 2019). According to previous research by Daubert et al. (2020); Jean et al. (2019), the use of pedagogical questions has created specific effects on the learning processes of young children. For example, Jean et al. (2019) notes that when attempting a complex task, children exposed to pedagogical questions perform a greater number of hypothesis tests, while Daubert et al. (2020)'s study revealed that books containing pedagogical questions improved children's psychosomatic understanding more than direct instruction or nothing at all.

Overall, psycholinguistic findings prompt us to ask what the role of question asking is not only in language acquisition, but also in training language models. We seek to explore the effect of questions on complex grammar understanding by artificially increasing the rate of questions in the data.

2.2 Language Models

Child-Directed Speech has been found to support language acquisition by better fitting the learner's needs in its unique linguistic and distributional properties (Nencheva and Lew-Williams, 2022; Eaves Jr et al., 2016). Following such findings, computational models have shown the advantages of using CDS as training data for Large Language Models (LLMs) in achieving similar performance with less data or better performance with the same amount of data (Eaves Jr et al., 2016; Huebner et al., 2021; Mueller and Linzen, 2023b; You et al., 2021). While these studies highlight the potential of CDS as training data, the amount of available CDS remains limited compared with the needs of most LLMs. For example, the NA-English portion of CHILDES (MacWhinney, 2000), the largest resource for CDS, amounts to 14.5M words in the 100M data release of the BabyLM challenge (Charpentier et al., 2025).

Hence, computational models have sought to artificially generate properties of CDS using non-CDS data, aiming at replicating CDS effectiveness. Huebner et al. (2021) has shown that using age-ordered CDS results in superior accuracy in learning the underlying grammar. To replicate the advantage of ordered input, curriculum learning models construct input streams from non-CDS data by gradually increasing complexity levels as measured in word diversity, abstractness, grammatical complexity, etc. (Tsvetkov et al., 2016). Since the BabyLM data (Jumelet et al., 2025) consists of both CDS and non-CDS data, several studies have applied curriculum learning to the data, showing

Dataset	Q%	Q-MLU	Yes/No%	Wh%	Examples
CHILDES	20.54	4.92	22.84	28.67	*"Is he gonna take a bath?"*, *"What color 's that?"*, *"yeah?"*
BNC	15.15	8.67	19.23	19.40	*"Doesn't he go out on Saturday night?"*, *"On the system?"*
Gutenberg	7.91	9.77	25.11	28.36	*"Is Lady Jane Ashleigh within?"*, *"What makes all these bushes grow here?"*
OpenSubtitles	17.74	5.38	17.52	31.04	*"Can I help you two?"*, *"Why did you break up?"*
Simple Wiki	0.08	11.71	2.30	25.29	*"What Ever Happened to Baby Jane?"*, *"London; a multicultural area?"*
Switchboard	4.05	6.92	29.44	19.49	*"How do you keep up with current events?"*, *"You're kidding?"*

Table 1: Statistical analysis of questions in each dataset: the percentage of questions out of all sentences, the MLU of questions in words, the percentage of Yes/No questions vs. Wh-questions, and examples of questions from each dataset.

some improvement, though models achieved notable performance gains by modifying the learning algorithm or adding new data (Hu et al., 2024).

A complementary approach aims to artificially create text data that either creates CDS-like textual content or augments non-CDS data to fit CDS characteristics. For example, Theodoropoulos et al. (2024) artificially created children's stories, which are known to provide enhanced learning opportunities for children (Montag et al., 2015, 2018). Haga et al. (2024) add variation sets to the BabyLM data by rephrasing sentences to create close sequences of semantic repetitions. While their results show mixed effects over different tasks, their analysis suggests that the prompting method used to create the variation sets might have reduced the word variability in a way that limits the performance.

Our approach is focused on one aspect - questions; however, we recognize that the resulting data may share some additional properties of CDS. The questions are generated as part of the turn-taking sequence of the existing data. As such, the artificial data adds semantically similar sentences into the sequence that can enhance learning. In addition, the questions may form a variation set by adding a question and an answer after an existing sentence with some overlapping words and concepts, as observed in variation sets. For example, the question *"Did he draw the sword from the stone?"* created by the model repeats the words *'sword'* and *'stone'* in a novel construction for this section of the dialogue.

3 Methods

We began with the data used for the 2025 BabyLM Challenge (Charpentier et al., 2025), provided by Jumelet et al. (2025), and used gpt-5-mini from OpenAI's API (OpenAI, 2025) to generate questions based on the content of the original files. We did not modify the data from CHILDES (MacWhinney, 2000) since we considered it to be the gold standard with regard to the ratio of questions to non-question statements, the type of questions, and their linguistic properties.[1]

Our data generation process is as follows:

Using the prompt: "You are a helpful reading companion for a 5 or 6-year-old child. Take the passage below. Ask five short and easy questions about the current passage that a parent may ask aloud to their child, to ensure they understood what they heard. After stating the question, exclaim the answer enthusiastically. Use child-directed speech: clear, friendly language, simple grammar. Focus on key details in the text (who, what, where, why, how – or yes/no). Keep each question under 10 words and end with a question mark. 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.", we asked gpt-5-mini(OpenAI, 2025) to generate questions based on the texts provided.

For each dataset, we combined the original data with the generated data without removing the ques-

[1]The data and models generated by our study can be found here: https://github.com/NLPlabMSU/BabyLM_Questions

tions already used in the data. To remain within the 10M-word limit, we randomly down-sampled each dataset so that the combined original and generated text was as close as possible to the original size. The resulting samples differed by no more than 10 words from the original data. This ensures that our total data size does not exceed that of the original. This also ensures that the proportion of file sizes between different files is the same.

Then, we trained multiple GPT-Wee (Bunzeck and Zarrieß, 2023) setups using the transformers (Wolf et al., 2020) package. We compared the following models: (1) a baseline model using the original data included in *Strict-Small*, (2) a model where every training file is replaced by its respective modified file with the questions, and (3) models where every training file except one is the original and only one modified file with questions. The third type results in five models, one for each dataset other than CHILDES(MacWhinney, 2000), which allows us to evaluate the contribution of augmenting each of the datasets to the overall performance.

To reduce the effect of variance, we ran five trials for each setup with different seeds, for a total of 35 models. The parameters for training are a batch size of 32, max steps of 40000 with evaluation every 10000 steps, 1000 warmup steps, 8 gradient accumulation steps, and a learning rate of 5e-4.

4 Results

The *Strict-Small* data consists of six data sets: CHILDES (MacWhinney, 2000), British National Corpus (BNC) (BNC Consortium, 2007), dialogue portion, Project Gutenberg (children's stories) (Gerlach and Font-Clos, 2020), OpenSubtitles (Lison and Tiedemann, 2016), Simple English Wikipedia (Wikimedia, 2023), and Switchboard Dialog Act Corpus (Stolcke et al., 2000). We consider CHILDES as the baseline for question asking in CDS and thus do not add questions to it or modify it in the simulation. We first present an analysis of the distributional properties of questions in CDS and each of the original datasets, and the augmented datasets. Second, we present the results using the original data vs. the augmented data to train our model.

4.1 Analysis of Question Distribution

Table 1 presents the distributional properties of questions appearing in each of the datasets in *Strict-Small*, including examples for each data. As expected, CHILDES has the highest percentage of questions in the data at 20.54%. Moreover, as expected from CDS, the Mean Length of Utterance (MLU) in words for CDS questions is the shortest with 4.92 words. We estimate the type of the question as a Yes/No question or Wh-question using the opening words of the questions. This method may overestimate the number of questions that are neither Yes/No nor Wh-questions since some questions start with a discourse marker or opening clause, e.g., *"Oh, are you?"*. However, this method follows psycholinguistic findings that emphasize the role of overlapping prefixes in aiding language acquisition as observed in child production (Cameron-Faulkner et al., 2003). We find that CDS contains 22.84% Yes/No questions and 28.67% Wh-questions.

We randomly sample the questions from each dataset under each question type category to illustrate the semantic and pragmatic properties of the questions. CDS contains many Yes/No questions relating to the semantic context of the question to verify understanding. Wh-questions can be seen used to prompt information seeking and extension. Finally, verification questions such as *"yeah?"* and *"she's poorly?"*. Table 1 provides additional examples from all datasets for the various question types. While we do not show the percentage of verification questions directly, many of the questions that are neither Yes/No nor Wh-questions fall under this question type.

OpenSubtitles has the closest percentage of questions to CDS (17.74%) and the closest MLU (5.38 words). However, this data has a much higher rate of Wh-questions over Yes/No questions and semantic scope that differs from CDS. The BNC dialogue portion follows the question rate of OpenSubtitles with 15.15%, but the MLU is higher than CDS with 8.67 words. The Gutenberg data shows the closest distribution of question types to CDS, which is consistent with its composition of children's books. However, the percentage of questions in the Gutenberg data is much lower than CHILDES, while the MLU is higher. Finally, both Simple English Wikipedia, and Switchboard Dialog Act Corpus have a relatively low rate of questions, which is expected from Wikipedia as a non-dialog source, but more surprising from Switchboard. Both datasets have higher MLU and semantic scope that cannot be matched with CDS.

Table 2 shows the same distributional properties for the artificial data. Notably, the dataset contains

Dataset	Q%	Q-MLU	Yes/No%	Wh%	Examples
BNC	22.21	10.76	24.74	25.55	*"Is this about angles and shapes?"*, *"Who is coming to stay?"*
Gutenberg	13.39	7.83	22.70	54.33	*"Who talked about Pink Pills?"*, *"Who came to help Bomba?"*
OpenSubtitles	18.64	6.02	18.48	35.42	*"Did he draw the sword from the stone?"*, *"Who was taken?"*
Simple Wiki	6.37	5.85	30.84	68.01	*"Was Nezval born in 1900?"*, *"Who became UN Secretary-General in 2017?"*
Switchboard	11.87	8.64	26.72	26.58	*"Who went with the kids to see different colleges?'*, *"'Did they talk about fly fishing?'*

Table 2: Statistical analysis of questions generated by our method: the percentage of questions out of all sentences, the MLU of questions in words, the percentage of Yes/No questions vs. Wh-questions in each dataset, and examples of questions from each dataset generated by the prompt.

both the original questions and those generated by our prompting method, as the original questions are retained rather than removed. While the rate of questions increases for all datasets, it remains lower or similar to the percentage of questions in CDS. The MLU of the questions varies from 10.76 to 5.85, but does not correlate with the MLU of questions or sentences in the original data. For example, the MLU of all sentences in the BNC dialogue portion is 10.80, and the artificial questions are of similar length. The MLU of all sentences in English Wikipedia is 12.61, but the MLU of the artificial question is only 5.85. We hypothesize that the MLU of the generated questions depends on the semantic properties of the data in addition to the syntactic ones. However, instead, it seems to be that topics where you can easily make questions with "correct answers" given common knowledge, like "who was George Washington", had lower Q-MLUs, whereas conversational corpora like BNC may result in longer questions since many questions would require context, i.e. "where did Mom go after picking the kids up from school". The artificial data also contains a much higher rate of Wh-questions, which could result from the prompt used to generate the data. We aim to explore prompting methods that elicit more verification questions in the future.

4.2 Learning from Question-Augmented Data

Our prompting method creates new questions based on text data that was already included in the baseline *Strict-Small* dataset. Thus, we do not predict significant changes in the semantic abilities and world-knowledge over the artificially generated data, though we hope to further explore these questions in the future. Instead, we focus our analysis on the BLiMP benchmark to evaluate how the artificial data affects the learning of particular areas of linguistic knowledge.

Table 3 presents the results for each of the sub-categories included in the BLiMP benchmark (Warstadt et al., 2020) and the overall average. We compare the results using the provided *Strict-Small* dataset (on the left), to the data generated by replacing all datasets with the same-size version with an increased rate of questions as explained in Section 3 (shown on the right side of the table). In the middle section of Table 3, we present the results for changing only one dataset at a time with its corresponding version with the increased rate of questions.

The overall performance on the BLiMP benchmark is better given the original data. Although the performance loss is low, it is consistent across simulations and also consistent with previous methods of artificial data augmentation such as that from the study by (Hu et al., 2024). While the overall performance was better with the original data, notably, all individual categories show the best performance for one of the models based on modifying only one of the datasets, or the addition of questions to all subsets. This result is somewhat surprising given the relatively low number of sentences introduced by each dataset individually. The positive impact of a single dataset modification confirms our hypothesis that questions can influence computational training similarly to their role in language acquisition.

The improvement to some categories can be at-

	10M	BNC	Gutenberg	OS	Wiki	Switchboard	10M-QA
Island Effects	41.39	41.14	40.76	40.17	41.48	40.30	**42.55**
Anaphor Agreement	75.97	75.51	73.21	**79.05**	75.53	75.26	70.76
Argument Structure	59.70	59.82	**60.58**	58.44	59.36	59.48	57.16
Determiner-noun Agr.	74.23	73.51	71.85	70.84	**74.28**	73.38	66.39
Subject-Verb Agr.	58.50	58.25	58.44	57.66	58.17	**58.70**	56.46
Ellipsis	57.55	57.32	54.36	57.39	57.97	**58.14**	54.54
Control/Raising	58.78	58.26	**59.80**	58.24	58.14	58.37	59.59
Quantifiers	83.23	82.11	81.60	82.17	78.06	**82.53**	67.70
Irregular Forms	82.64	**83.56**	75.81	82.96	82.10	82.17	74.40
NPI Licensing	54.35	**54.95**	54.03	52.17	51.73	53.67	54.52
Binding	64.85	65.20	64.23	64.06	**66.23**	65.19	64.15
Filler Gap	65.24	65.17	65.82	64.82	65.51	65.39	**66.06**
Average	**62.13**	62.01	61.45	61.61	61.91	61.91	59.49

Table 3: Averaged scores in % trained for 40000 steps. Models differ only in training data: (1) 10M Original - *Strict-Small* data provided by BabyLM, (2)-(6) 10M original with one dataset switched with a version enhanced with questions and answers, and (7) all datasets replaced with the versions enhanced with artificial questions and answers. The top score for every category is marked in bold.

tributed to the type of linguistic challenge captured by the task. For example, as expected, we observe a positive impact on the performance for Island Effects and Filler Gap categories. These results align with the high rate of Wh-questions generated by the prompting method. Moreover, the category of improvement can be analyzed with respect to the linguistic properties of the datasets before the modification and the behavior of the question-generation method for this dataset. English Simple Wikipedia and Switchboard had the lowest percentage of questions in the original data and a relatively high MLU. The modified data for these datasets result in performance gains for Determiner-Noun Agr. and Binding (English Wikipedia) and Subject-Verb Agr. and Ellipsis categories (Switchboard).

The modified data for the BNC dataset results in better performance on the Irregular Forms category. The modified data for the Gutenberg datasets improves the results for Argument Structure and Control/Rising categories. Finally, OpenSubtitles modification results in better performance on Anaphor Agreement. We hypothesize that each of these results can be explained by considering the linguistic properties of the specific data set. For example, the Gutenberg data consists of children's stories, a type of data that has been suggested to play an important role in argument structure learning (Montag et al., 2015, 2018).

Interestingly, in several categories, the addition of questions to each subset results in improvement to the score, while adding questions to all the datasets results in a significant drop. For example, 10M-QA scores for Quantifiers and Irregular Forms are 67.70% and 74.40%, while the top score for each is 82.53% and 83.56% respectively. These differences lead to the overall lower score for the 10M-QA compared with the baseline, despite the benefit of adding questions to each dataset. We hypothesize that the disadvantage of adding all questions relates to the difference in the linguistic properties of the questions in CDS vs. the synthetic data. We discuss future directions to extend our analysis in the next section.

5 Discussion

Child-Directed Speech differs from Adult-Directed Speech in many ways. It has been shown to better support both language acquisition and computational modeling. Due to the limited availability of CDS compared with other datasets, the ability to generate CDS-like data using AI-generated text can improve training ability. In this study, we focused on the increased rate of questions in CDS as a possible linguistic characteristic that may support learning. Our results show a positive effect only for directly related grammatical categories, e.g., Island Effects and Filler Gap. Moreover, our analysis of the data generation shows a potential sensitivity to the linguistic properties of the data used for prompting over the prompt itself in guiding the model on the target generative goal.

Contrary to our predictions that the datasets with

questions would perform better, our results actually demonstrated that questions lowered overall BLiMP (Warstadt et al., 2020) performance, with the models where every training file had been replaced with the enhanced questions data performed the worst, while the original performed the best. It should be noted that the best performance for most BLiMP task categories resulted from the addition of questions to one of the datasets. However, none of the datasets consistently improved all tasks compared to others or to the baseline. Furthermore, this pattern seems to be supported by the fact that the models in which the augmented file is relatively small–such as Switchboard (Stolcke et al., 2000)–performed better than those in which the augmented file is larger, such as Gutenberg (Gerlach and Font-Clos, 2020).

To further understand the results, we analyzed the fine-grained performance on all subtasks included in the BLiMP benchmark. All trained models performed generally poorly at determining when to use "that" vs. a Wh-word when the verb of which it is an object is far away, but not when it is directly connected to the verb of interest. For example, the baseline model averaged around a 7.62% accuracy score on the wh_vs_that_with_gap_long_distance benchmark, but a 98.68% accuracy with the wh_vs_that_no_gap_long_distance benchmark (Warstadt et al., 2020). The length of the sentence or the clause does not seem to affect this very much as the models performed about equally well on the wh_vs_that_no_gap benchmark and the wh_vs_that_no_gap_long_distance benchmark (Warstadt et al., 2020).

One potential cause of the overall degradation in performance when questions are introduced to the training data could be that AI-generated questions may not be the same as the kind of pedagogical questions and verification questions asked by parents or educators in child-directed speech. Another possibility is that a high portion of our questions fell under the Wh-questions category, which may have reduced the grammatical diversity of the data overall when modifying all datasets. We observe that many questions in CDS do not take the syntactic form of questions, but rather rephrase previous content as a declarative sentence for verification or clarification. Thus, while CDS offers diverse training data, the syntactic questions may cause a bias in the distribution over syntactic forms that prevents the model from learning all grammatical categories adequately.

Importantly, although we prompted the model with the same instructions for all datasets, including the limitation on question length and complexity, the model failed to produce consistent linguistic properties for all questions across all datasets. This unexpected behavior might be advantageous for linguistic diversity overall. Language learning relies on exposure to both simple and complex argument structure, so the ability of the generative model to adapt to the linguistic properties of the input might be to the benefit of the downstream training. To fully explore this question, we aim to analyze the linguistic diversity of the generated data as well as the model's performance on additional benchmarks.

A high percentage of the questions in CHILDES and other conversational datasets included verification questions that do not fall under either Yes/No or Wh-questions. These questions offer continuous semantic context while diversifying the argument structure and choice of words, as they often repeat recent communication for verification goals. The use of verification questions is tightly connected to the use of variation sets and close repetitions in CDS. We aim to explore alternative prompts that emphasize the use of verification questions in addition to other types of questions in the future by considering alternative prompts. We also hope to extend our analysis to annotate the questions with their communicative goal, e.g., pedagogical questions, to better understand the generation of artificial data and its effect on model training. This study shows the potential in adding questions to datasets in order to enhance the learning of certain linguistic properties. This preliminary study offers quantitative and qualitative analysis, which offers multiple directions for future research and linguistic exploration.

6 Limitations

We used GPT-5-mini (OpenAI, 2025) in order to generate questions for the texts. Some attempts to, for example, create CDS based on the OpenSubtitles file were thwarted by the model's guardrails due to the violent/explicit content of the movie subtitles being deemed inappropriate for children. We overcame this behavior by adding to the prompt that it was "for research purposes". Likewise, early attempts at prompting the model generated formatted text with emojis and other extraneous charac-

ters, thus we directly addressed this by expanding our prompt to exclude those characters.

Due to computational resources and space limitations, we cannot detail the full scope of experimented prompts. Some outputs were somewhat nonsensical or tangential to our request given minimal trials. Also, the output length of the GPT-5-mini model made it impossible to pass the model the entire training file, so it was split into chunks. However, even those reasonable-sized chunks were too large and required to generate a separate file containing questions and add them back into the training files. In the future, as LLMs' context windows expand, we may be able to more efficiently explore further.

Another limitation we faced was regarding computing power. Affordable compute power and GPU access are often limited. We were only able to run ten trials per setup with the resources we had, but we aim to extend this analysis in the future.

Acknowledgments

We thank Rachel Hamelburg and Aliyah Vanterpool for providing us code and assistance in designing the prompting process. We thank Dr. Feldman, Dr. Peng and the NLP lab at Montclair State University for helpful discussion and feedback.

References

Igor Bascandziev, Patrick Shafto, and Elizabeth Bonawitz. 2021. The sound of pedagogical questions. In *Proceedings of the annual meeting of the cognitive science society*, volume 43.

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.

Maureen A Callanan and Lisa M Oakes. 1992. Preschoolers' questions and parents' explanations: Causal thinking in everyday activity. *Cognitive development*, 7(2):213–233.

Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. 2003. A construction based analysis of child directed speech. *Cognitive science*, 27(6):843–873.

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. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Emily N Daubert, Yue Yu, Milagros Grados, Patrick Shafto, and Elizabeth Bonawitz. 2020. Pedagogical questions promote causal learning in preschoolers. *Scientific reports*, 10(1):20700.

Baxter S Eaves Jr, Naomi H Feldman, Thomas L Griffiths, and Patrick Shafto. 2016. Infant-directed speech is consistent with teaching. *Psychological review*, 123(6):758.

Lukas Edman and Lisa Bylinina. 2023. Too much information: Keeping training simple for BabyLMs. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 89–97, Singapore. 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, Leshem Choshen, Ryan Cotterell, Alex Warstadt, and Ethan Gotlieb Wilcox, editors. 2024. *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Miami, FL, USA.

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.

Anishka Jean, Emily Daubert, Yue Yu, Patrick Shafto, and Elizabeth Bonawitz. 2019. Pedagogical questions empower exploration. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 41.

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

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.

Jessica L Montag, Michael N Jones, and Linda B Smith. 2015. The words children hear: Picture books and the statistics for language learning. *Psychological science*, 26(9):1489–1496.

Jessica L Montag, Michael N Jones, and Linda B Smith. 2018. Quantity and diversity: Simulating early word learning environments. *Cognitive science*, 42:375–412.

Aaron Mueller and Tal Linzen. 2023a. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. *arXiv preprint arXiv:2305.19905*.

Aaron Mueller and Tal Linzen. 2023b. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

Mira L Nencheva and Casey Lew-Williams. 2022. Understanding why infant-directed speech supports learning: A dynamic attention perspective. *Developmental Review*, 66:101047.

Elissa L Newport, Henry Gleitman, and Lila R Gleitman. 2020. Mother, i'd rather do it myself. *Sentence first, arguments afterward: Essays in language and learning*, 141.

OpenAI. 2025. Gpt-5 mini. https://openai.com. Lightweight variant of GPT-5 large language model.

Meredith L Rowe. 2008. Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of child language*, 35(1):185–205.

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

Catherine E. Snow and Charles A. Ferguson, editors. 1977. *Talking to Children: Language Input and Acquisition*. Cambridge University Press.

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 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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Guanghao You, Balthasar Bickel, Moritz M Daum, and Sabine Stoll. 2021. Child-directed speech is optimized for syntax-free semantic inference. *Scientific Reports*, 11(1):16527.

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