Using Curriculum Masking Based on Child Language Development to Train a Large Language Model with Limited Training Data

Evan Lucas[1], Dylan Gaines[1], Tagore Rao Kosireddy[1], Kevin Li[1], Timothy C. Havens[1]

[1]Michigan Technological University
1400 Townsend Drive
Houghton, Michigan, United States

Abstract

In this paper we detail our submissions to the STRICT and STRICT-SMALL tracks of the 2024 BabyLM Challenge. We approach this challenge with two methodologies: i) use of a novel dataset, and ii) development of a pre-training technique based on the fusion of child language acquisition with traditional masked language modeling, which we call *curriculum masking*. The novel dataset used for this task is based on user submissions to the Reddit forum (i.e., subreddit) "Explain Like I'm Five", which explains diverse concepts using simple language. Curriculum masking works by creating learning phases based on a standard child language development timeline, where the masked words learned by the model start with simple nouns and gradually expand to include more complex parts of speech. We show that using internet-based training data shows a small improvement in evaluation scores as compared to baseline training data. Our proposed pre-training method of curriculum masking is conceptually novel and also shows improved rates of learning over typical masked language modeling pre-training, potentially allowing for good performance with fewer total epochs on smaller training datasets. Code for the curriculum masking implementation is shared at https://github.com/evanperson/curriculumMaskingBabyLM2024.[1]

1 Introduction

Children acquire language skills through exposure to an estimated two to seven million words per year. However, contemporary large language models (LLMs) require training on massive datasets comprising billions to trillions of words to achieve similar linguistic capabilities. The vast disparity between human language acquisition and current machine learning practices can be shown from the Chinchilla model (Hoffmann et al., 2022), which was trained on 1.4 trillion words.

To address these disparities, the BabyLM Challenge was established to explore the feasibility of pre-training LLMs on datasets comparable in size to those encountered during early childhood language development. It continues on this mission, imposing strict limits on the size and composition of training datasets and aims to create models that learn language in a child-like manner.

In this paper, we present our submissions to both the STRICT and STRICT-SMALL of the 2024 BabyLM Challenge. We leverage a novel dataset, sourced from the Reddit forum (i.e. subreddit) *Explain Like I'm Five* (ELI5). We introduce a curriculum masking training strategy that we designed to mimic how children learn language. Traditional *Masked Language Modeling* (MLM) masks random words during training, which does not accurately reflect the structured manner in which children acquire language. Our curriculum masking approach organizes the process into a schedule of stages, starting with simpler words such as nouns, then gradually incorporating more complex words like adjectives and verbs. We hypothesized this scheduling method would help the model build a stronger foundation in language before tackling more advanced sentence structures. We do not count the added information of POS tags as additional word count, as it is only an additional categorical variable attached each token. In conversations on the challenge Slack channel, this was agreed to not count towards overall word count for this reason, and so although it is additional information, we make the claim that it does not count as increased training words. The following sections of this paper describe our dataset preparation, implementation of curriculum masking, experimental results, and discussion of the effectiveness of our approach.

[1]https://github.com/evanperson/curriculumMaskingBabyLM2024

2 Related Work

To understand the state of related work, we reviewed the most relevant papers from the BabyLM 2023 challenge (Warstadt et al., 2023) and performed searches for similar concepts to what we propose in this work. Several works in the prior challenge utilized curriculum learning, though they administered their curricula through an intentional sequencing of the examples in their training set as opposed to the curriculum-based masking approach we use.

Martinez et al. (2023) investigated curriculum learning strategies for language model pre-training using limited data. Similar to our curriculum masking, their approach progressively increased task complexity, structuring model training in stages. Although their methods did not consistently outperform non-curriculum baselines, their focus on vocabulary and data pacing offers valuable insights for optimizing training with limited resources.

DeBenedetto (2023) applied a curriculum learning strategy for low-resource settings, where datasets were ranked by difficulty using a bytes-per-line metric. Simpler datasets, such as spoken transcriptions, were introduced first, then more complex datasets were gradually introduced during training. Their approach outperformed baseline models in most downstream tasks, including BLiMP and SuperGLUE. Their curriculum learning methods demonstrated consistent improvements in performance, particularly when trained with more epochs.

Bunzeck and Zarrieß (2023) designed a curriculum learning approach based on child-directed speech which showed an improvement for certain tasks like anaphor agreement, irregular forms, and quantifiers. Their work, similar to ours, involved using curriculum learning focused on word frequency and sentence structure. However, they used a static data ordering approach where the training data was organized in a fixed sequence.

Curriculum masking as a concept has been applied successfully in other domains, such as computer vision. Jarca et al. (2024) developed a curriculum based masking strategy for vision tasks. They show that by using a curriculum-based masking approach for training vision models they are able to outperform the same model architecture on some common image classification tasks.

3 Method

For our submission, we make two primary contributions: i) a new ELI5 dataset, and ii) a method of curriculum learning that modifies the MLM pretraining task to mimic child language development. In this section we review these two contributions as well as the other method choices made.

3.1 Dataset

The curated dataset provided by the challenge organizers contains various sources of child and child-directed speech. In addition to this dataset, we created a novel dataset using the subreddit *Explain Like I'm Five*[2]. On this particular subreddit, users can pose questions on almost any topic. Other users' responses to these questions are required to be free of technical jargon and tend to use simplified concepts. While the responses are not targeted at actual five-year-olds, we felt the nature of ELI5 could be a good fit for the BabyLM Challenge. We chose not to use the existing ELI5 dataset (Fan et al., 2019) as it is focused more on question answering and we needed text for pretraining that met our dataset objectives.

We obtained all of the posts to the ELI5 subreddit from June 2005 to December 2022 from The-Eye.eu Reddit archive[3]. We filtered this set of posts to leave only top-level comments, which are direct replies to questions, by searching for posts where the link ID matched the parent ID in the metadata. Only top-level comments are required by the subreddit rules to be simplified explanations. We then sorted the remaining posts in descending order based on the score they obtained through the built-in user voting system on the subreddit. In the event two posts had the same score, the more recent post came first in the sorted list. We applied a basic filter that removed posts containing any of the profane words in a 28-word list. We also removed posts that included "https" to filter out hyperlinks from our training data. This profanity filter is fairly simple and it is not likely that it removed all instances of profanity within the dataset (e.g. alternatively spelled profanity). However, profanity is a part of language and even children are exposed to a non-zero amount of profane language during their developmental years. Thus, we did not conduct further filtering beyond the simple list.

We created training sets of 10M and 100M

[2]www.reddit.com/r/explainlikeimfive

[3]https://the-eye.eu/redarcs/

words. We computed the number of words in each post by splitting the text on any white space character and summing the number of text segments that contained any alphanumeric character. We developed this method to count towards the limit any word that contains meaning. It does not count punctuation that stands alone, such as dashes. Working down the sorted list of posts, we added posts to each training set as long as it did not cause the sum to exceed the total word capacity. Due to this, our 10M word training set is a proper subset of the 100M word training set.

3.2 Curriculum Development

Classic masked language model training involves randomly masking tokens from the training data that is fed to the model. With a limited amount of training data, we sought to develop a curriculum that more closely mimicked how a child might learn language. Using a timeline for normal child language development (LaGreca, n.d.; Roseberry-McKibbin and Hegde, 2006), we developed the following steps:

1. Interjections, nouns, and personal pronouns
2. Conjunctions
3. Subject-verb-object structures, first person singular pronouns, plurals, and simple verb forms
4. Adjectives, plural proper nouns, possessives, wh-determiners, and pronouns
5. Complex verbs and possessive endings
6. Adverbs, particles, and complex adjectives
7. All other parts of speech

The curriculum was cumulative, so each step in the training contained the additional parts of speech for that step and all previous steps.

We used the Natural Language Toolkit (NLTK) library (Hardeniya et al., 2016) to implement part-of-speech (POS) tagging on our datasets. We converted each sentence in our training data into individual words and obtained the POS tag for each word from the toolkit. We then created a custom function that selectively masked words based on their grammatical categories. For each training example, we drew masked words from the pool of tags for the current curriculum step until either 15% of the total words were masked or all candidate words were used.

Table 1: Hyperparameters used for training

	100M	10M orig.	10M redo
Learn. rate	5e-5	5e-5	1e-4
Optimizer	AdamW	AdamW	AdamW
LR Profile	Linear	Linear	Cosine
Warmup	n/a	n/a	500

3.3 Base model selection and computing parameters

We used a BERT (Devlin, 2018) model with 6 layers, a hidden dimension of 768, and 12 heads to create a 51.2M parameter model. Following RoBERTa (Liu, 2019) and GPT-2 (Radford et al., 2019), we used a *byte pair encoding* (BPE) tokenizer to tokenize the inputs. We built tokenizers with 10K and 50K vocabularies for each of the 10M and 100M corpora, respectively. We used two sets of hyperparameters: the original set used for initial model training and an updated set that we used for a final training pass. These hyperparameters are compiled in Table 1. These parameters were largely arbitrarily chosen based on past experience by the authors and we note that there are probably additional changes to these design choices that could be made to improve performance.

We used a variety of GPUs and workstations to train and evaluate our models, including six 40Gb A100s, an A6000, two RTX2080Tis, and two RTX3090 GPUs. We estimate our combined GPU-days at around 30 days. Due to the varying VRAM available on each of these GPUs, batch sizes were not consistent between the training of different models and we note this as a weakness in our study. Past experience from one of the authors has shown that batch size is a particularly important parameter for small datasets as a bigger batch size smooths the loss landscape and reduces the capacity of the model to learn from individual examples. Private conversations with some industry members have suggested that in very small datasets, it's sometimes desirable to fine-tune with a batch size of one in order to learn the distribution of the data. However, due to the time constraints of this challenge, we maximized batch size to make use of the available GPUs and did not well-control or study it.

4 Results

In this section, we share the results of our models on the BabyLM shared task. We have attempted

Table 2: **Evaluation scores** The overall evaluation metrics we computed for all trained models. We did not train or evaluate models with a '_provided' suffix and the results presented come from the challenge organizers. The first BERT_10M_eli5_curr run results files do not have available GLUE scores and therefore no macroaverage.

Model	BLiMP	BLiMP Supplement	EWoK	GLUE	Macroaverage
BERT_10m_base	54.6	56.5	47.3	66.1	56.1
BERT_10m_eli5	54.7	56.5	49.9	66.8	57.0
BERT_10m_eli5_curr_mask_redo	55.6	56.1	**50.8**	**67.3**	57.5
BERT_10m_eli5_curr_mask_orig	51.25	52.23	48.0	xx.x	xx.x
LTG-BERT_10M_provided	**60.6**	**60.8**	48.9	60.3	**57.7**
BERT_100m_eli5	55.4	54.0	51.5	66.7	56.9
BERT_100m_eli5_curr_mask	60.2	56.8	**53.0**	67.7	59.4
LTG-BERT_100M_provided	**69.2**	**66.5**	51.2	**68.4**	**63.8**

to disentangle the impacts of the two approaches combined, although due to training time we were not able to do a full ablation study. First, we discuss the impact of the newly scraped dataset. Second, we share the results of the curriculum masking approach and discuss why it appears to outperform the typical MLM pre-training approach.

4.1 ELI5 dataset

Findings by Meta (Xie et al., 2024) show that having a high fraction of internet scraped data is generally the key to the highest performing language models. We decided that we would try to go with a solely internet-based training dataset in an attempt to take advantage of this effect. Following anecdotes from the training of the original Stable LM (Bellagente et al., 2024), which used a high fraction of Reddit-based training data and had poor performance, our data cleaning removed usernames (suspected to be responsible for strange tokenizer performance in Stable LM and GPT-2). To be able to identify the impact of our ELI5 dataset, we trained an identical model on the baseline dataset provided by the BabyLM organizers. In Table 2, the *BERT_10m_base* and *BERT_10m_eli5* entries show the baseline and ELI5 data evaluation results, respectively. LTG-BERT scores provided by the organizers were included as a fair comparison for the encoder-only BERT model we used. BLiMP scores show barely any difference, suggesting that grammatical phenomena are represented similarly in both datasets. EWoK, a benchmark evaluating world knowledge, shows improved results with the ELI5 dataset, which the authors find to be a reasonable outcome due to the simplistic explanations that capture world knowledge found in many ELI5 responses. SuperGLUE evaluations also show modest improvements from use of the ELI5 dataset, potentially indicating that the ELI5 data teachesa better language understanding than the baseline training dataset. Comparing both to the LTG-BERT results provided by the competition organizers (*LTG-BERT_10M_provided*), the BLiMP results of both BERT models are lower, but all other metrics are higher for our models.

When looking at the 100M models, they appear to underperform the provided model's results on BLiMP, be somewhat comparable on GLUE, and greatly outperform the provided model on EWoK. We did not spend as much time adjusting hyperparameters for or rerunning the 100M data, so it is likely there is a lot of room for improvement. However, despite this, these results reinforce our finding that ELI5 explanations help teach world knowledge to a language model.

4.2 Curriculum Masking

As described in Section 3.2, the curriculum masking process gradually introduced the model to new parts of speech, while continuing to train on the previously introduced parts of speech. Training results from the improved training hyperparameters (*BERT_10m_eli5_curr_mask_redo* in Table 2) (the hyperparameters listed in Table 1) are shown in Figure 1. The learning rates follow a cosine decay with a warmup period that resets every time a new part of speech is introduced. Loss shows small increases with the introduction of new parts of speech and then gradually decays as is typically expected. The gradient norm increases as the parts of speech become more complex, indicating that the model is not learning as well with

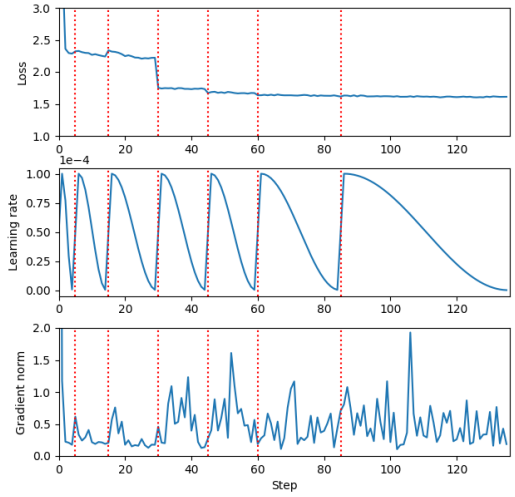

Figure 1: Curriculum masking training performance for 10M ELI5 training.

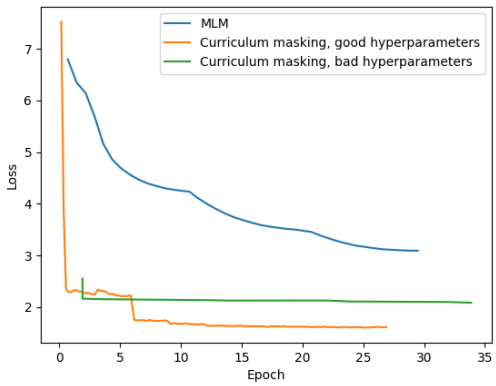

Figure 2: Loss curve comparison for curriculum-based and traditional masked language modeling with 10M model.

a very limited set of parts of speech but during the intermediate and later stages of the curriculum it learns quite effectively. By comparing with two other loss curves in Figure 2, we demonstrate that this loss curve outperforms either a linearly decaying learning rate that resets with each new POS or a typical masked language modeling approach with linearly decaying learning rate with three restarts (*BERT_10m_eli5_curr_mask_redo* and *BERT_10m_base*, respectively, in Table 2). We note that this is not a strong comparison as the logging rates do not match due to the batch size mismatch, but the general trends may be helpful for the reader. Importantly, we note that by focusing the model on learning specific aspects of language first, we are able to accelerate the learning of the more complex language aspects introduced later.

When looking at evaluation scores in Table 2, with the better hyperparameters, we demonstrate that the combination of the ELI5 data with the curriculum masking provides the best performance overall of any 10M model we evaluated. We note that BLiMP performance was comparatively poor for all of the models we trained, relative to the provided scores of the baseline model. For the 100M models, the curriculum masking improved results beyond just using the ELI5 dataset, although with poor BLiMP performance, the improvements to EWoK weren't able to increase the average score above the baseline model results provided by organizers.

5 Conclusions

In our submission to the 2024 BabyLM challenge, we focus on using an internet-based training dataset that mimics language that would be directed at youth as well as utilizing a developmentally plausible pre-training approach that allows the model to learn specific parts of speech on a schedule. We show that by combining both of these approaches, we can outperform the baseline provided by organizers on the STRICT-SMALL track (10M word limit) of the challenge, although we did not succeed at outperforming the baseline for the SMALL track (100M word limit). Due to a lack of hyperparameter optimization, there is probably a lot of improvement that could be made using curriculum masking, especially considering different masking ratios or masking ratio schedules. One other possibility we are interested in is varying the POS acquisition order and experimenting with the use of training the model on a mix of the POS on the schedule as well as some other words. Testing the curriculum masking concept on an autoregressive model would be an obvious thing to try as well.

Our findings help reinforce the idea that using internet-scraped data provides highly useful data for teaching a language model language understanding as well as world knowledge. Additionally, our proposed method of curriculum masking introduces a new method of curriculum learning that shows accelerated learning in our tests on a small dataset.

Limitations

The biggest limitation of this work is that it largely relies on two sets of hyperparameters and does not thoroughly explore the hyperparameter space in order to determine how stable and useful our proposed training method is. Masking rates were not explored at all and there are most likely masking rates or schedules for them that would further improve model training and performance. We have tried to explore the impact of our data and training method separately by running a partial ablation study, but we did not consider the impacts of data on our hyperparameter selection. Batch size is often noted as a powerful "knob" for tuning performance and due to the mismatched GPUs used for different training runs, we did not control this well and therefore are not able to quantify its impact on our model performance. There is also a dependence on the performance of the POS tagger and we don't have a good assessment of the performance of the POS tagger used without having labeled data from our dataset.

Ethics Statement

The authors are not anticipating any major ethical concerns with publishing this work. We propose a slight modification to the widely-used MLM pretraining task as well as a version of a publicly available dataset. We note that our use of the ELI5 Reddit data encourages the continued use of scraped internet data to train language models, which has been noted to potentially lead to self-training on generated content as more internet content becomes generated by language models. The long term impacts of this are not fully understood yet, but it is likely that it may be somewhat detrimental to both future model performance and, thus, internet content.

Acknowledgments

This work was supported by the Michigan Tech Institute of Computing and Cybersystems and by grant funding received from NIH/NIDCD R01DC009834.

References

Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable lm 2 1.6 b technical report. *arXiv preprint arXiv:2402.17834*.

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.

Justin DeBenedetto. 2023. Byte-ranked curriculum learning for babylm strict-small shared task 2023. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 198–206.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Nitin Hardeniya, Jacob Perkins, Deepti Chopra, Nisheeth Joshi, and Iti Mathur. 2016. *Natural language processing: python and NLTK*. Packt Publishing Ltd.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Andrei Jarca, Florinel-Alin Croitoru, and Radu Tudor Ionescu. 2024. Cbm: Curriculum by masking. *arXiv preprint arXiv:2407.05193*.

Lauren LaGreca. n.d. Normal language development for young children. https://www.lispeech.com/normal-language-development-young-children/. Accessed: 2024-09-19.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Richard Diehl Martinez, Zebulon Goriely, Hope McGovern, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. Climb: Curriculum learning for infant-inspired model building. *arXiv preprint arXiv:2311.08886*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Celeste Roseberry-McKibbin and Mahabalagiri N Hegde. 2006. *An advanced review of speech-language pathology: preparation for PRAXIS and comprehensive examination*. ERIC.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal

Linzen, et al. 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*.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2024. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36.

A Dataset examples

A few semi-cherry picked examples are shown for some of the ELI5 data and some of the provided baseline training data. It can be seen that the internet-based text of ELI5 is more coherent and provides a better textual training example (in the subjective opinion of the authors) than the transcribed text that is formatted in a variety of ways. Whether or not an explanation given at a level appropriate for a five year old is equivalent to what a five year old actually experiences is debatable, but from the language modeling perspective it is likely that the transcribed text may cause the model to learn strange behaviors that are not reflective of actual language usage.

ELI5 Sample response 1

The joke answer is so that the water doesn't hit you square in the face.

The real answer is that shapes with sharp corners are structurally weak. Arcs and circles are very strong shapes. If port holes were squares, the openings would get damaged and worn out sooner.

ELI5 Sample response 2

Caffeine works in two ways to make you feel that way.

First it prevents the brain from telling you that you are tired. You can think of your brain as a bunch of locked boxes with different things inside of them. Some of these boxes have things that make you happy, others make you sad. Some have things that tell you it is time to go to sleep. Caffeine jams itself into the lock on the sleepy time box so that your brain can't open it. That keeps you from feeling tired.

Caffeine also can help open the box that tells your body to go into extra energy mode. Things like your heart can work faster or slower depending on what you need. If you are sitting on the couch watching TV it's going to go slower, if you are outside working it's going to speed up. Caffeine tricks the body into thinking it needs to go into extra energy mode. Caffine doesn't create this energy, the body is just using what it has stored more quickly. Not really any different from you step on the gas in a car. You are telling it to burn more fuel and go faster.

ELI5 Sample response 3

You know when you're going on vacation, and you're packing, but you still need to use some of the stuff you need to pack, so instead of putting it all into your suitcase, you set some of it next to your suitcase, or leave it out on the counter, so you don't forget it, but you can still use it without having to completely unpack it from your luggage?

That's sort of how a USB drive works. Sometimes you tell the computer to "pack" data onto the drive, and rather than put it all on there right away, it might end up caching some of it to be written later.

When you just rip out the drive, you risk pulling it before all of your data is "packed" onto the drive.

When you click "safely remove" it runs around the house and packs up all the stuff it left out, and gets it all into the luggage for you before you disconnect it.

BabyLM Provided 10M_Train Sample 1

Have you ever seen anybody completely obscured by her own smoke, it's Sharon.
.
Chuck us the water would you?
She's a bit of a goer as well int she?
Is she?
Isn't she?
Yeah but
Didn't she order a punch so she was drunk ?
No, that was Tracey.
I thought Tracey and Sharon used to get drunk at lunchtime on a Friday and have a punch up.
No.
Only Tracey would do that.
Our Trace.
Ah.
Oh dear.
Oh.

BabyLM Provided 10M_Train Sample 2

THIS IS EXACTLY WHAT I'M TALKING ABOUT.
I'M NOTHING BUT A BIG MAC IN A BATH TOWEL.
JOEY, I'M NOT A HAMBURGER.
I HAPPEN TO BE A HUMAN BEING.
JESS, BUDDY, AS LONG AS I'M THE DIRECTOR,
YOU WILL BE TREATED WITH DIGNITY AND RESPECT.
THANK YOU.
OK, HOSE HIM DOWN.

BabyLM Provided 10M_Train Sample 3

*CHI: Eve tapioca hot.
MOT: uhhuh.
CHI: hot.
MOT: mhm.
CHI: and cool.
MOT: and cool yes.
MOT: by the time you have lunch it'll be cool.
CHI: that?
MOT: what is that?
MOT: vanilla.
CHI: vanilla.
MOT: vanilla.
CHI: vanilla.
MOT: vanilla.
CHI: Eve play bouillon cube.