Is Child-Directed Speech Effective Training Data for Language Models?

Anonymous ACL submission

Abstract

While high-performing language models are typically trained on hundreds of billions of words, human children become fluent language users with a much smaller amount of data. What are the features of the data they receive, and how do these features support language modeling objectives? To investigate this question, we train GPT-2 models on 29M words of English-language child-directed speech and a new matched, synthetic dataset (TinyDialogues), comparing to a heterogeneous blend of datasets from the BabyLM challenge. We evaluate both the syntactic and semantic knowledge of these models using developmentally-inspired evaluations. Through pretraining experiments, we test whether the global developmental ordering or the local discourse ordering of children's training data support high performance relative to other datasets. The local properties of the data affect model results, but somewhat surprisingly, global properties do not. Further, child language input is not uniquely valuable for training language models. These findings support the hypothesis that, rather than proceeding from better data, children's learning is instead substantially more efficient than current language modeling techniques.

1 Introduction

Transformer-based language models (LM) show very strong performance on a wide variety of downstream tasks, but typically only after pretraining on hundreds of billions to trillions of words (Brown et al., 2020). In contrast, human learners use language fluently after far less training data – in the 10s to 100s of millions of words. This "data gap" (Frank, 2023a) of several orders of magnitude poses a substantial challenge for machine learning.

Is the source of human children's efficient learning a function of their data or their learning algorithms? While children receive rich multi-modal input from their exploration of the world, here we focus primarily on their language input, which has been a major focus of study in developmental psychology (MacWhinney, 2014). One hypothesis is that the language data that children receive is a uniquely rich learning signal – conversational interaction with their caregivers – that is curricularized optimally to support learning (Eaves Jr et al., 2016; You et al., 2021; Newport, 1990). Indeed, interventions to increase the quality of caregiver language do produce improvements in children's language learning (Ferjan Ramírez et al., 2020), and interventions to simplify model training data also result in stronger performance (Muckatira et al., 2024; Eldan and Li, 2023).

Language model pretraining experiments provide a targeted method for investigating dataset quality (Kallini et al., 2024): we can manipulate the training data available to models to create "controlled rearing" experiments. We take advantage of this method to investigate the properties of child-directed speech for learning the syntactic and semantic structure of language. We use GPT-2 as our simulated learner, and pretrain on natural and synthetic child-language data. For each of these, we conduct two experiments. First, we investigate whether the natural curricularization of children's input – from simpler utterances to more complex conversations – affects language model learning. Second, we test whether the local discourse coherence structure of dialogue results in better learning. Finally, we compare to learning on a more heterogeneous blend of data from various sources.

We find that the curricularization of child language does not provide a uniquely valuable signal for language models, supporting the hypothesis that other aspects of children's learning (not simply the data) – perhaps interactions with their training data – are responsible for their efficiency relative to language models. On the other hand, the source, composition, and local properties of the training data have measurable effects on model performance.

2 Related Work

The efficiency of children's learning has been an important focal point for recent NLP efforts (Huebner et al., 2021; Zhang et al., 2021). For example, last year's BabyLM challenge held the training data for models constant, while encouraging entrants to investigate alternative learning architectures (Warstadt et al., 2023). Smaller models of this type must be evaluated using targeted benchmarks more appropriate to their performance levels, including evaluations of semantic (Zhuang et al., 2023) and grammatical abilities (Huebner et al., 2021; Warstadt et al., 2020). These evaluations have even been used to benchmark performance based on data from a single child (Qin et al., 2024).

The method of "controlled rearing" (manipulating data while holding the model constant) for language models (Frank, 2023b) has a long history in cognitive science, e.g. Christiansen and Chater (1999), but has recently become prominent for testing learnability claims (Warstadt and Bowman, 2024; Kallini et al., 2024; Misra and Mahowald, 2024). Often, models trained on naturally-occurring corpora are contrasted with counterfactual corpora constructed via targeted experimental manipulations – for example, shuffling sentence ordering (Kallini et al., 2024) or removing particular constructions (Misra and Mahowald, 2024).

Curricularization of training data is widely investigated in machine learning (Bengio et al., 2009), with the guiding idea being that an appropriate ordering of training examples can lead to a smoother path to the desired objective. Children's development is argued to create a curriculum to facilitate their learning (Smith et al., 2018; Cusack et al., 2024). In one study from the visual domain, Sheybani et al. (2024) trained self-supervised models on data from infants and found that a developmental ordering leads to stronger eventual performance compared with a reversed ordering. Our study tests this hypothesis in the language domain.

3 Methods

3.1 Datasets

CHILDES The Child Language Data Exchange System (CHILDES) is a repository of human-transcribed corpora of children and caregivers' talk (MacWhinney, 2014), with children ranging from birth to age 13. We take the English subset, which consists of approximately 29M total words (including speaker labels and other metadata) across ≈11k conversations. CHILDES is heavily skewed towards younger ages; ≈90% of the data is for children ages 2-5 (see Figure 1 in Appendix C).

TinyDialogues Inspired by TinyStories (Eldan and Li, 2023), we collect a synthetic dataset consisting of approximately 29M words called TinyDialogues (TD). Using GPT-4, we prompted the generation of realistic conversations involving children of ages 2, 5, 10, and 15 years as the central participant, along with a list of other potential participants (e.g. mom, teacher, babysitter). To diversify, we seeded each conversation based on a list of words known by children at the relevant age and varied the conversation type and length (see Appendix A).

BabyLM We further compare to the dataset distributed by the BabyLM challenge (Warstadt et al., 2023), a 100M word dataset that is a mixture of several sources including transcribed speech, child-directed speech (e.g. CHILDES), children's storybooks, and simple Wikipedia. It is designed to approximate the language data that a 10-year-old child could receive. We sub-sampled ≈29M words from BabyLM to match the size of this dataset to our CHILDES and TinyDialogues data.

Preprocessing Training data for CHILDES and TD was set up so that each line corresponded to a single conversation. Training data for BabyLM was set up using the pre-existing format in the BabyLM challenge, which varied depending on the sub-dataset. Each dataset was then split into 85/15 train/val splits, of approximately 24.5M training words and 4.5M validation words. During training, GPT-2 breaks up the data into chunks of 1024 consecutive tokens which are fed into the model.

3.2 Evaluation

Zorro (Huebner et al., 2021) is designed for child-directed language and aims to quantify the syntactic and grammatical knowledge of language models. It does so by assessing their capability to distinguish between minimal pairs of sentences that exhibit various grammatical contrasts. We report final averages (of accuracy, higher is better) across individual Zorro tasks in Section 4.

Word Similarity To assess the semantic knowledge of our models, we employ a word similarity (WS) metric (Zhuang et al., 2023), which measures the ability of models to capture semantic similarities between pairs of words. Many assessments of children's language learning are multi-modal and involve visual stimuli; in contrast, word similarity is a language-only measurement that can be adapted for assessing LMs. We extract word embedding representations from hidden layers of each model, compute pairwise cosine similarities between these embeddings, and report Spearman correlations between human and model similarity judgments (higher is better). The best layer of each model is chosen. We average results across several word similarity benchmarks including RG-65 (Rubenstein and Goodenough, 1965), WordSim-353 (Finkelstein et al., 2001), SimLex-999 (Hill et al., 2015), SimVerb-3500 (Gerz et al., 2016), and MEN (MTest-3000) (Bruni et al., 2012).

3.3 Experiments

Global Ordering To test whether the natural ordering of speech to children presents an effective curriculum for model learning, we ordered our CHILDES and TD conversations (training examples) in three ways: 1) age order (from younger to older), 2) reverse order (from older to younger), and 3) random order (equivalent to default LM training practices of randomly shuffling the training data). CHILDES includes fine-grained age information of the target (main) child involved in each conversation, down to fractions of months (essentially days), and we ordered conversations based on this information. TD was ordered based on the conversation seed ages of 2, 5, 10, and 15 years old. For the random order experiments, we randomly shuffled the conversations and kept this shuffled order for all experiments for consistency purposes.

Local Ordering To investigate the effects of local conversation and discourse coherence on learning, we ordered utterances within each CHILDES and TD conversation in two ways: 1) normal (original) order, and 2) random order. The latter breaks the local discourse coherence, by randomly shuffling utterances within each conversation.

3.4 Model Training

We use GPT-2 (Radford et al., 2019) with 124M parameters (small version), following prior "controlled rearing" work (Kallini et al., 2024; Misra and Mahowald, 2024; Qin et al., 2024). We trained a separate tokenizer on each of our datasets, and pretrained GPT-2 from scratch using a learning rate (LR) of $1e-04$, linear LR scheduler with no warmup steps, varying batch sizes (e.g. 4, 8, 16) per GPU, training seed of 42, and Adam optimizer with $\beta = (0.9, 0.999)$ and $\epsilon = 1e-08$.

Model	Zorro	WS
CHILDES	77.77%	0.24
TD	79.42%	0.41
BabyLM	81.75%	0.42

Table 1: Evaluation results of our GPT-2 models across datasets, using standard iterative training for 20 epochs (used to train BabyLM) and random global ordering.

Model	Zorro	WS
CHILDES	74.24%	0.18
TD	77.05%	0.32

Table 2: Evaluation results of our GPT-2 models, comparing natural (CHILDES) vs. synthetic (TD) conversation data. These results are averaged across all relevant experiments: different global and local ordering methods and curricularization strategies.

For our global ordering experiments, we split each dataset into b approximately equal sections (buckets), and trained on each repeatedly (n times) before moving to the next bucket or section of the dataset. This technique was intended as a compromise between standard techniques for model training – which require iterated training on a dataset – and human learning – which operates via a single pass through ordered training data. In Section 4, we report results averaging across $n = 3, 5, 10, 20$. For TD, we used the data corresponding to the four seed ages as the four *buckets*. For CHILDES, we experimented with different numbers of buckets (b) and settled on $b = 5$ for most experiments. To compare to BabyLM (which cannot be *bucketed*), we also trained iteratively on each dataset for 20 epochs, selecting the epoch that performed best on the respective validation split (lowest val loss).

4 Results and Analysis

Major results of our experiments can be found in Tables 1, 2, 3, and 4. More detailed results of the curricularization experiments (i.e. broken down by different values of b and n) can be found in Appendix G.

As seen in Tables 1 and 2, training on BabyLM yields noticeably better results than CHILDES and TD on Zorro (syntax), and substantially better results than CHILDES on WS (semantics). Overall, these results are surprising: a mixture of different data sources may be more effective for training

Model	**Order**	**Zorro**	**WS**
CHILDES	Age	73.54%	0.19
CHILDES	Reverse	74.18%	0.18
CHILDES	Random	74.76%	0.18
TD	Age	77.24%	0.31
TD	Reverse	75.77%	0.31
TD	Random	77.54%	0.32

Table 3: Evaluation results of our GPT-2 models, comparing global ordering methods, broken down by dataset. These results are averaged across relevant experiments (different training and curricularization methods).

Model	**Order**	**Zorro**	**WS**
CHILDES	Normal	77.65%	0.21
CHILDES	Random	77.23%	0.16
TD	Normal	80.67%	0.37
TD	Random	79.03%	0.37

Table 4: Evaluation results of our GPT-2 models, comparing local ordering methods, broken down by dataset. These results are averaged across two experiments each: standard 20-epoch training and repeated buckets ($b = 5, n = 10$) using random global ordering.

smaller models on a limited amount of data, and even synthetic conversation data appears to be more effective than natural conversation data for training small-scale models on limited data.

Some potential explanations include fundamental differences in the data itself. CHILDES is heavily skewed towards younger ages (see Figure 1 in Appendix C), whereas TD is a more uniform distribution across ages with more sophisticated conversations intended to simulate speech to older children. As such, it contains a higher fraction of more grammatical utterances and text. While collecting TD, we ensured that TD was also diverse in the type of conversation, participants, and content. This feature likely leads to a more comprehensive coverage of the distribution of potential conversations. Both these factors may lead to more effective learning of syntax and semantics. Lastly, high-quality synthetic data – in contrast to naturalistic data, which contains disfluencies and occasional garbled tokens due to transcription issues – may simply be better suited for training LMs, especially when data is limited (and quality is likely even more important).

As seen in Table 3, the effects of global ordering seem to have negligible effects on model performance. Zorro and WS values remain relatively stable despite changes to global order. This result is also surprising, as one would expect that, similar to humans, curricularization based on difficulty would affect model learning and performance, i.e. it would be easier to learn from simpler utterances and conversations before moving to more difficult ones. Aligning with this, model convergence behavior, especially training loss, differed on a local level (e.g. per epoch or bucket) depending on the ordering method, but general high-level behavior (especially of the validation loss) was relatively stable (see Appendix I). Language models, unlike humans, may be less affected by curricularization, especially when the amount of data is limited.

From Table 4, we see that local ordering affects model performance. Breaking local discourse coherence negatively impacts both Zorro and WS evaluations, even though Zorro consists of single sentence evaluations. Effects on WS are more evident for CHILDES training than TD, where WS remains the same. A potential explanation is that CHILDES contains noticeably shorter utterances, on average, than TD ($\approx$ 4 words vs. 13). Hence, reordering CHILDES utterances likely has a greater effect on the model's ability to learn semantics from similarity across a larger set of short utterances.

5 Conclusion & Future Work

Why do children require so much less training data than language models to become fluent language users? We ran experiments using GPT-2 on CHILDES, BabyLM, and a new synthetic conversation dataset that we collected called TinyDialogues. A heterogeneous mixture of different data sources performed better than homogeneous conversation data – further, high-quality synthetic conversation data yielded better performance than natural conversation data. Additionally, we found that global developmental ordering and curricularization did not have noticeable effects on performance, whereas local discourse coherence structure did. In sum, it seems that the curricularization of child language does not provide a uniquely valuable signal for language models. However, the source, composition, and local properties of the training data affect model learning. We hope that future work builds on our work here to expand upon the available evaluation benchmarks and data mixtures for comparison between models and children, and extends this comparison to multi-modal datasets and models.

Limitations

Some limitations of our work include our current suite of evaluation benchmarks and models. We can expand our benchmarks to include more theory of mind and developmental psychology-inspired benchmarks, and ones for longer coherency evaluation. We can also experiment with larger language models such as LLama-3. Furthermore, we limited our investigations to conversation data and the BabyLM mixture. We could explore more types and sources of data, and different varieties and proportions of data mixtures. Additionally, the CHILDES dataset is heavily skewed towards younger ages. Unfortunately, to the best of our knowledge, a more balanced and uniform dataset of high-quality textual transcriptions of child-directed conversations is not currently available, but we could consider collecting one in the future. Overall, these are directions to potentially improve and expand upon our work in the future. We feel that, despite these limitations, our current work is an insightful and focused contribution.

Ethical Considerations

The majority of our datasets and evaluation benchmarks are already existing, publicly available datasets and benchmarks, intended for public use.

We collected TinyDialogues using GPT-4, following all intended use purposes and OpenAI's policies. Further, the dataset is entirely synthetic, and does not include personal or private information. As a safe and controlled language model, there is an incredibly low risk of offensive content, especially as it involves conversations with younger children. We also manually examined a large subset of the data and ensured there were no ethical issues. This includes profanities, racism, bias, offensive words, and other malicious language.

We acknowledge the potential weaknesses of our trained models, which are small in scale and limited in performance. We will never use or encourage their use for real-world purposes. Our initial experiments are conducted purely for investigation purposes to test our hypotheses. We feel that our work is an important contribution to the ML, NLP, cognitive science, and psychology communities, and we encourage researchers to expand upon it.

Our models, TinyDialogue dataset, and accompanying publication are intended only for research purposes and to assess the effectiveness of child-directed speech for training language models. We do not foresee any explicit way that malicious actors could specifically misuse our trained models or models that could be trained on our dataset.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145, Jeju Island, Korea. Association for Computational Linguistics.

Morten H Christiansen and Nick Chater. 1999. Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, 23(2):157–205.

Rhodri Cusack, Marc'Aurelio Ranzato, and Christine J Charvet. 2024. Helpless infants are learning a foundation model. *Trends in Cognitive Sciences*.

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.

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

Naja Ferjan Ramírez, Sarah Roseberry Lytle, and Patricia K Kuhl. 2020. Parent coaching increases conversational turns and advances infant language development. *Proceedings of the National Academy of Sciences*, 117(7):3484–3491.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. *ACM Transactions on Information Systems - TOIS*, 20:406–414.

Michael C Frank. 2023a. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Michael C Frank. 2023b. Openly accessible llms can help us to understand human cognition. *Nature Human Behaviour*, 7(11):1825–1827.

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. 2021. *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

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.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible language models. *arXiv preprint arXiv:2401.06416*.

Victor Kuperman, Hans Stadthagen-González, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior Research Methods*, 44:978–990.

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

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing aanns. *Preprint*, arXiv:2403.19827.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent abilities in reduced-scale generative language models. *Preprint*, arXiv:2404.02204.

Elissa L Newport. 1990. Maturational constraints on language learning. *Cognitive science*, 14(1):11–28.

Yulu Qin, Wentao Wang, and Brenden M. Lake. 2024. A systematic investigation of learnability from single child linguistic input. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*.

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.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. In *Communications of the ACM (CACM) 8 (10)*.

Saber Sheybani, Himanshu Hansaria, Justin Wood, Linda Smith, and Zoran Tiganj. 2024. Curriculum learning with infant egocentric videos. *Advances in Neural Information Processing Systems*, 36.

Linda B Smith, Swapnaa Jayaraman, Elizabeth Clerkin, and Chen Yu. 2018. The developing infant creates a curriculum for statistical learning. *Trends in cognitive sciences*, 22(4):325–336.

Alex Warstadt and Samuel R. Bowman. 2024. What artificial neural networks can tell us about human language acquisition. *Preprint*, arXiv:2208.07998.

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*.

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.

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.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Chengxu Zhuang, Evelina Fedorenko, and Jacob Andreas. 2023. Visual grounding helps learn word meanings in low-data regimes. *arXiv preprint arXiv:2310.13257*.

A TinyDialogues: Dataset Collection Details & Examples

Here we discuss some further dataset collection details for TinyDialogues (TD), with examples of TD conversations in Table 5.

The specific GPT-4 model we use for collecting our entire dataset is *gpt-4-1106-preview*, which is GPT-4 Turbo with training data up to Apr 2023. To increase the diversity of the generated conversations, when prompting GPT-4, we also specify the particular type of conversation (Table 6), the approximate length or number of turns (5 or 10),[1] other potential participants in the conversation (Table 7), and certain words (one noun, one verb, and one adjective) sampled from Wordbank CDI (Frank et al., 2021) (ages 2 & 5) and AoA (Kuperman

[1]GPT-4 had a tendency to generate longer conversations, around 10 and 20 turns instead, respectively.

et al., 2012) (ages 10 & 15), cut off by the seeded age, that must be included in the conversation for content diversity. The list of potential participants and content words varied by age, e.g. a 15-year-old teenager would likely not talk regularly with a babysitter. We also collect some additional metadata: a list and description of all participants in the conversation, and a brief description of the context/setting. We only use the dialogue portions for our experiments. The GPT-4 prompt is below.

GPT-4 Prompt: *Please construct a realistic, approximately {5, 10}-turn dialogue directly involving a {2, 5, 10, 15}-year-old {toddler, child, teenager}*[2] *as a participant. The {toddler, child, teenager} is the central participant in the dialogue, with most/all speech directed towards them. Hence, for this dialogue, please limit the vocabulary to that of which a typical {2, 5, 10, 15}-year-old {toddler, child, teenager} would understand. The dialogue should be {type}.*[3] *The dialogue should use the verb '{verb}', the noun '{noun}', and the adjective '{adjective}'. Please include the following participants along with the child: {participants}.*[4] *Participant labels should be surrounded by double asterisks, i.e. '**participant**'. If there are several of the same type of participant (e.g. multiple friends or classmates), please label them distinctly, e.g. '**Friend 1**' and '**Friend 2**'. Please list and describe the participants after 'PARTICIPANTS:', briefly describe the context/setting of the dialogue after 'SETTING:', and present the dialogue itself after 'DIALOGUE:'. The turns of the dialogue should be separated by '*`\n\n`*'. Remember, please ensure the dialogue is realistic, and one that would likely occur in the real world directly involving a {2, 5, 10, 15}-year-old {toddler, child, teenager}."*

B Data Format & Preprocessing

CHILDES & TD: We noticed some duplicate utterances in CHILDES conversations and removed these to the best of our ability. See below for an example of part of a single training example for CHILDES. Double asterisks surround speaker labels, double newline tokens separate utterances, and an end-of-text token marks the end of the conversation. This format was consistent across all conversations in both CHILDES and TD datasets.

***CHI**: Are those your stars?* `\n\n` ***MOT**: Can you say star?* `\n\n` ***CHI**: Star.* `\n\n` ***CHI**: Look.* `\n\n` ***CHI**: Stars.* `\n\n` ***MOT**: Stars. See? Look, look at the yellow star, a golden star. <|endoftext|>*

BabyLM: We concatenated the data across the BabyLM sub-datasets, then sampled approximately 29M words to match the amount of data in CHILDES and TD, while keeping the original distribution among its sub-datasets intact. We sampled in order (i.e. starting from the beginning of each sub-dataset), as several BabyLM examples (e.g. for Wikipedia) span multiple lines, and random sampling would have broken up individual examples and eliminated coherence. We do no further preprocessing to the BabyLM data and keep the format of the original sub-datasets. In particular, we do not add an <|endoftext|> token to the end of each example (like we do with CHILDES and TD) as it is unclear where each example ends.

The Python NLTK package's *word_tokenize* function was used as part of our statistics calculations (discussed in Appendix C). The parameters for this function are: *language* is 'english' (default) and *preserve_line* is 'False' (default) so line breaks are ignored. Specifically, it was used for calculating the number of unique words in Appendix C. For consistency purposes, data processing and sampling, and other word-related statistics (e.g. total word count, avg. words per utterance) were done by simply splitting the text by space.

C Dataset Statistics

CHILDES consists of ≈11k conversations over ≈29M words. The mean length of utterances is low, at only 3.85 words (minus speaker label), which is likely partially due to the skew in age, where ≈90% of the data is for children ages 2-5 (see Figure 1). CHILDES contains ≈74.5k unique words and ≈448 utterances (on avg.) per conversation.

BabyLM consists of ≈443k unique words in the ≈29M word subsample we take for our experiments. Individual example statistics are not available as some examples (e.g. Wikipedia) span multiple lines, and no end of example markers were given. More specific details and statistics about the dataset (including its sub-datasets) can be found in Warstadt et al. (2023), e.g. Table 1 in their paper.

TinyDialogues consists of ≈130k conversations

[2]*toddler* is used for age 2, *child* for ages 5 and 10, and *teenager* for age 15.

[3]A random conversation type along with its explanation is sampled each time from the ones in Table 6.

[4]If $turn = 5$, we randomly sample one additional participant from the corresponding list in Table 7. For $turn = 10$, we randomly sample two additional participants.

Age	Example
2	**Babysitter**: Look, the pool is all empty. All gone! \n\n **Toddler**: Water all gone? \n\n **Babysitter**: Yes, all gone. We let the water out. It went down, down, down. \n\n **Toddler**: Why? \n\n **Babysitter**: So the pool can dry. We don't want it to get yucky. \n\n **Toddler**: Pool fall? \n\n **Babysitter**: No, the pool didn't fall. The water fell out when we opened the plug. \n\n **Toddler**: Plug? \n\n **Babysitter**: Yes, the plug is like a little stopper that keeps the water in. When we pull it, water comes out. \n\n **Toddler**: More water? \n\n **Babysitter**: Not now. We'll fill the pool again another day, okay? \n\n **Toddler**: Okay. Play ball? \n\n **Babysitter**: Sure, let's play ball now. The pool will be here when we're ready to fill it up again.
5	**Teacher**: "Alright, everyone, it's time to clean up! **Child**, can you please help me by putting the crayons back in the box?" \n\n **Child**: "Yes! I can do that. The box is empty so I'll fill it up!" \n\n **Teacher**: "Thank you, that's very helpful. Make sure the lids are on tight so they don't dry out." \n\n **Child**: "I did it! Look, they're all inside now." \n\n **Teacher**: "Great job! Now, **Classmate**, can you and **Child** go and put the books back on the shelf, please?" \n\n **Classmate**: "Can we go to play after?" \n\n **Teacher**: "Yes, as soon as our classroom is tidy, we can all go outside and play." \n\n **Child**: "Come on, let's put the books away fast so we can go play!" \n\n **Classmate**: "Okay, let's do it! Race you to the shelf!" \n\n **Teacher**: "Be careful running, but I love the enthusiasm! Thank you both for helping."
10	**Dad**: "Once upon a time, in a faraway kingdom, there lived an earless rabbit who loved to make pancakes." \n\n **Child**: "An earless rabbit? How could he hear if he wanted to flip the batter?" \n\n **Dad**: "Well, you see, this rabbit had a special talent. He could feel the vibrations of the batter sizzling on the pan. When it was time to flip, he'd give it a perfect toss." \n\n **Child**: "That's so cool! Did the rabbit have any friends?" \n\n **Dad**: "Yes! His best friend was a turtle who loved to swim. One day, they decided to have a pancake picnic by the lake." \n\n **Child**: "Did they swim in the lake after eating pancakes?" \n\n **Dad**: "They sure did. The turtle taught the rabbit how to swim, and they had the best day splashing around until the sun set."
15	**Girlfriend**: Hey, so what's the plan for this history project video? \n\n **Teenager**: We need to make a mini-documentary about the industrial revolution. I was thinking we could start by showing how machines changed production, like how they used to churn butter by hand before. \n\n **Girlfriend**: Oh, cool idea! We could use that old butter churn in your grandma's attic for a visual. What role do you want me to play in the video? \n\n **Teenager**: Could you narrate the parts about the technological advancements? You're really good at explaining stuff. \n\n **Younger Sibling**: Can I help too? I want to be in the video! \n\n **Teenager**: Sure, you can help us set up the scenes. But no forcible taking over, okay? We need to work together as a team. \n\n **Younger Sibling**: I promise I'll be good! Can I churn the butter for the scene? \n\n **Teenager**: That's perfect! It'll look more authentic with you doing it. Let's get everything ready and start filming. Thanks for helping out, both of you.

Table 5: Examples of collected TinyDialogues conversations by seed age.

Conversation Type	Explanation
Explanatory	It should involve explaining something(s) and potentially answering question(s).
Functional	It should involve attempting to get something(s) done or accomplishing particular goal(s).
Narrative	It should involve telling a story (real or fictional) or sharing/recounting an experience.
Argumentative	It should involve conflict(s) or disagreement(s) that lead to an argument. In most cases, the argument should be resolved, resulting in the {child, toddler, teenager} learning.

Table 6: The four TinyDialogues conversation types along with their explanations.

across ≈29M words. There are ≈14 utterances (on avg.) per conversation, ≈96k unique words, and 13.42 words (on avg.) per utterance (minus speaker label). Since TD contains a uniform distribution across ages (including older ages), it is not surprising that the word diversity and average length of utterance are higher than CHILDES. Further, the average TD conversation is shorter than CHILDES, resulting in a higher number of individual conversations. More detailed statistics for TD (broken down by age) are in Table 8. As expected, the vocabulary (unique words) and average length of utterance increase with age. Conversely, the total number of conversations and average utterances per conversation decrease with age.

D Further Training & Compute Details

We searched through different values of the learning rate (LR) for GPT-2 training. Specifically, $LR = \{1e-06, 5e-06, 1e-05, 5e-05, 1e-04, 5e-04, 1e-03\}$. Through initial experiments, we found that $LR = 1e-04$ seemed to result in the best convergence behavior across the board, and used that for all our training experiments.

Our experiments were run on varying GPUs. This included a single RTX 3090TI (24GB VRAM), up to eight A40s (48GB VRAM each), up to eight A100s (80GB VRAM each), and up to four

TD Seed Age	Potential Participants
2	{mom, dad, older sibling, babysitter}
5	{mom, dad, older sibling, younger sibling, teacher, friend, classmate, grandparent, babysitter, neighbor}
10	{mom, dad, older sibling, younger sibling, teacher, friend, classmate, grandparent, babysitter, neighbor}
15	{mom, dad, older sibling, younger sibling, teacher, friend, classmate, grandparent, neighbor, coach, tutor, boyfriend, girlfriend}

Table 7: The list of other potential participants in each TinyDialogues conversation by seed age.

Age	Conversations	Total Words	Unique Words	Avg. Utterances per Convo	Avg. Words per Utterance
2	43,697	7,183,704	16,269	15.75	8.32
5	33,248	7,194,902	15,534	14.01	13.36
10	27,198	7,196,356	42,508	13.61	17.35
15	25,589	7,199,752	73,484	12.88	19.77
Total	129,732	28,774,714	95,868	14.29	13.42

Table 8: TinyDialogues dataset statistics broken down by seed age.

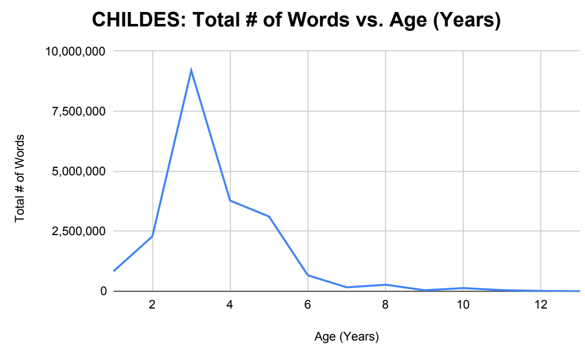

Figure 1: Total CHILDES word counts (utterances only, no metadata) by age.

H100s (80GB VRAM each). Training time varied by the type and number of GPUs and the particular experiment (e.g. number of repeated buckets).

E Zorro Evaluation Details

Zorro was inspired by BLiMP (Warstadt et al., 2020) and is a modification for child-directed language (e.g. lower vocabulary). However, it was designed specifically for masked language models such as RoBERTa. To adapt it to GPT-2, we reformatted the Zorro data to match the BLiMP format and used the BLiMP evaluation in the BabyLM evaluation suite[5] since the difference is mainly just the data. Further, we use the full Zorro test suite and do not filter examples by vocabulary. Hence, our results are not comparable to Qin et al. (2024) who filter Zorro examples by training vocabulary.

To better match the training data format and assess the effects of speaker labels, we came up with three variations of Zorro: 1) the original Zorro sentences (used to assess BabyLM), 2) the sentences with a common CHILDES speaker label prepended (used to assess CHILDES), and 3) the sentences with a common TD speaker label prepended (used to assess TD). To further match the training data as shown in Appendix B, the speaker labels were surrounded by double asterisks, and sentences included double newline tokens (before and after). We used the mother speaker label (MOT) for CHILDES, and the child speaker label (Child) for TD (see paragraph below), as these were some of the most frequent speaker labels for each dataset respectively (see Table 9). Further, preliminary experiments showed that these particular labels worked best for assessing each model.

One inconsistency in the collected TD data was that the speaker label for the target child varied across conversations and ages. For 2-year-olds, GPT-4 labeled the child *toddler*, and 15-year-olds were labeled *teenager*. This is likely due to our prompt as discussed in Appendix A. Further, within the same age, sometimes the label also differed (e.g. *Child*, *5-year-old child*, *5-year-old*). To align with CHILDES, which only used the speaker label CHI for every target child, and make Zorro evaluation consistent across the board, we converted most plausible target child speaker labels in our TD dataset (based on manual examination) to *Child*.

[5] https://github.com/babylm/evaluation-pipeline-2023

Dataset	Speaker Label	Frequency	Proportion
CHILDES	MOT	1,905,187	45.7%
CHILDES	CHI	1,593,073	38.2%
CHILDES	INV	188,712	4.5%
CHILDES	FAT	164,248	3.9%
TD	Child	735,176	46.6%
TD	Mom	132,746	8.4%
TD	Dad	129,568	8.2%
TD	Older Sibling	120,468	7.6%

Table 9: List of the top speaker labels for CHILDES and TD training splits along with their frequencies and proportions. This is after converting all target child labels for TD to *Child*, as described in Appendix E. For CHILDES: MOT stands for *mother*, CHI for *child*, INV for *investigator*, and FAT for *father*.

F Further Experimental Motivation

If a dataset can be described as a concatenation of equal-sized buckets A, B, and C, the repeated bucket approach can be described as $An\ Bn\ Cn$. The inspiration for this approach is due to several reasons, other than being a compromise between standard iterated training and human learning (as discussed in Section 3.4). Firstly, the iterative approach (training across the entire dataset several times) can potentially *wash away* global ordering effects (especially when the epoch count is high) as global order differences mainly exist within each individual epoch. When trained across several epochs, its effects may be less noticeable. The repeated buckets approach maintains the global order across training as a whole. The model can learn more from each bucket before moving to the next.

G Further Experimental Results

In addition to the experiments discussed in Section 4, we tried different values of n (number of times to repeat each bucket) for CHILDES and TD repeated buckets experiments. In particular, $n = 3, 5, 10, 20$. The chosen models for the repeated bucket experiments are the final models at the end of training. For CHILDES, we also tried different values of b (number of buckets, or approximately equal sections to divide the dataset into) using the global age order. In particular, $b = 3, 5, 10$. We report average results for different values of n and b in Tables 10 and 11, respectively. We also compare the typical iterative training approach (20 epochs) to repeated buckets using $n = 20$ (analogous to 20 epochs). Results are in Table 12.

Model	n	Zorro	WS
CHILDES	3	68.89%	0.10
CHILDES	5	72.02%	0.14
CHILDES	10	77.01%	0.19
CHILDES	20	75.75%	0.23
TD	3	71.51%	0.18
TD	5	74.48%	0.23
TD	10	79.21%	0.32
TD	20	79.65%	0.41

Table 10: Evaluation results of our GPT-2 models, comparing different values of n, broken down by dataset. These results are averaged across three different global ordering methods: age order, reverse order, and random order. For CHILDES, we use $b = 5$.

Model	b	Zorro	WS
CHILDES	3	73.36%	0.35
CHILDES	5	72.12%	0.35
CHILDES	10	70.06%	0.35

Table 11: Evaluation results of our CHILDES GPT-2 models, comparing different values of b. These results are averaged across three experiments each: global age order with $n = 3, 5, 10$.

H Importance of Speaker Labels

As an additional experiment, we also assess the importance of having speaker labels for each utterance. We train some versions of our models after removing all speaker labels (including their surrounding double asterisks). The results are reported in Table 13. As seen, removing speaker labels greatly detriments syntax and grammar learning (Zorro), but semantics (WS) appears unaffected.

I Convergence Behavior of Models

We plot the convergence graphs (train and validation losses vs. epoch) for several sets of our experiments in Figures 2 to 8. For the repeated buckets experiments, we treat the entire training run as a single epoch. We can notice interesting patterns/trends in the convergence behavior of models depending on several factors including the global ordering and curricularization method.

From Figure 2, we see that BabyLM converges to higher losses than CHILDES and TD, although it seems to perform better at test-time for syntax and semantics (as discussed in Section 4). Losses during training could simply be higher as the dataset is more complicated and varied since it is a mixture.

From Figure 3, we can see that when we train using the typical iterative epochs approach, the train-

Model	Approach	Zorro	WS
CHILDES	20-epochs	77.13%	0.52
CHILDES	$b = 5, n = 20$	75.75%	0.48
TD	20-epochs	79.41%	0.54
TD	$n = 20$	79.65%	0.54

Table 12: Evaluation results of our GPT-2 models, comparing typical iterative training (20 epochs) vs. repeated buckets with $n = 20$ (and $b = 5$ for CHILDES), broken down by dataset. These results are averaged across three experiments each: the three different global ordering methods (age order, reverse order, random order).

Model	Speaker Label?	Zorro	WS
CHILDES	Yes	77.65%	0.21
CHILDES	No	71.84%	0.22
TD	Yes	80.67%	0.37
TD	No	76.53%	0.37

Table 13: Evaluation results of our GPT-2 models, comparing speaker label vs. no speaker label for conversation utterances, broken down by dataset. These results are averaged across two experiments each: global random order with the typical iterative approach (20 epochs) and $n = 10$ (with $b = 5$ for CHILDES).

ing loss has a cyclical pattern using global age order and reverse order, while it converges smoothly for random order. From Figures 4 and 5, we see that when using the repeated buckets approach for both CHILDES and TD, global age order leads to a slowly cyclical increase in the training loss, while it generally decreases for reverse and random order. Throughout these experiments, while the training loss differs and individual buckets exhibit differing patterns, the high-level behavior of the validation loss and hence overall learning is similar. This aligns with the results we saw in Section 4.

From Figures 6 and 7, we see that varying b and n result in minor changes in behavior for the training loss. Specifically, by increasing n, the training loss has a more clearly defined cyclical pattern, and the losses converge to lower values. This is expected, since increasing n is analogous to training on more epochs. From Figure 8, we see that local interventions – randomly shuffling utterances and removing speaker labels (see Appendix H) – have minor effects on convergence behavior.

J Licenses and Intended Use

We used all existing datasets and models for their intended use. GPT-2 is licensed under the Modified MIT License. CHILDES is made available under TalkBank which is governed by the Creative Commons CC BY-NC-SA 3.0 copyright license (see https://talkbank.org/share/rules.html). We plan to release the TinyDialogues dataset under the standard MIT license. Information about the BabyLM challenge and its dataset (which is a collection of portions of several sub-datasets) is at https://babylm.github.io/index.html.

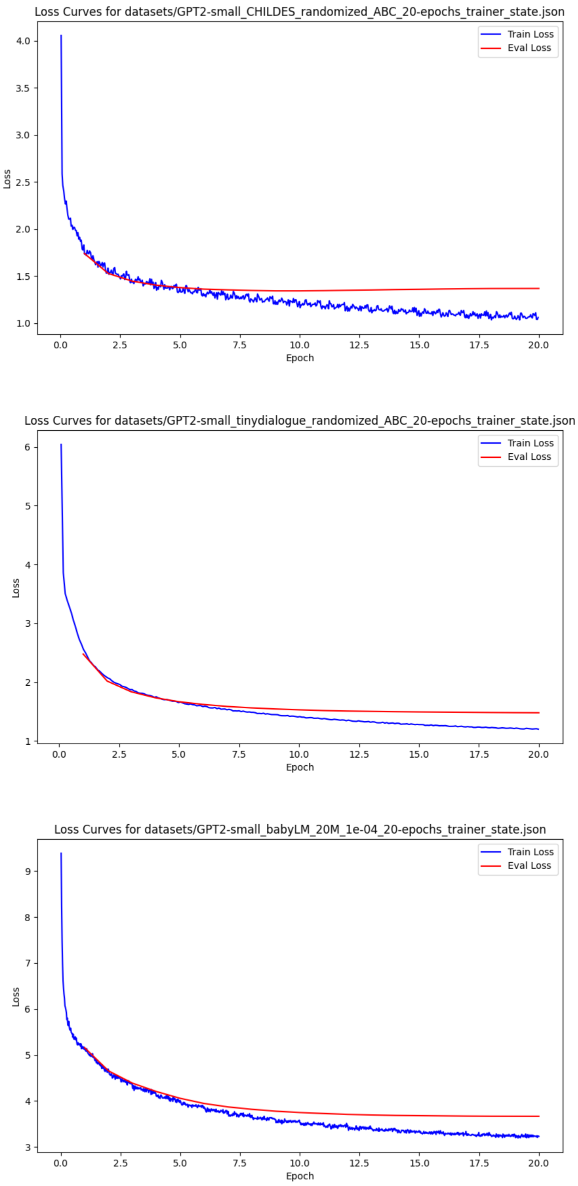

Figure 2: Convergence graphs (train and val loss) by dataset, using iterative training for 20 epochs. From top to bottom: CHILDES, TinyDialogues, BabyLM.

K Code & Data

We plan to publicly release all our code and data. Some of the code was written with the assistance of ChatGPT (specifically, GPT-4 and GPT-4o).

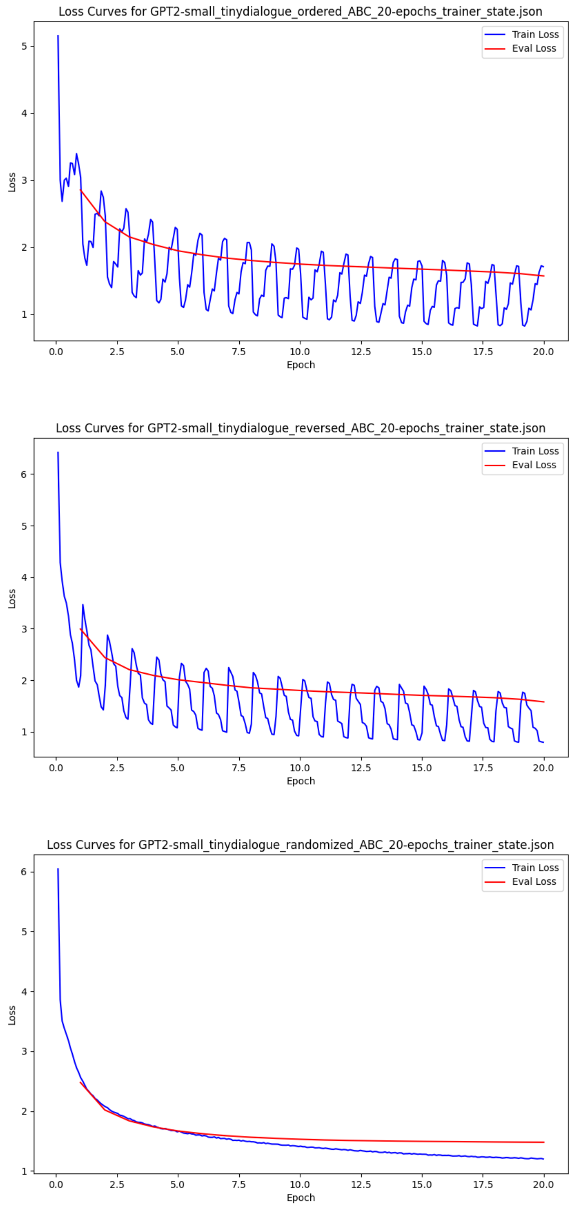

Figure 3: Convergence graphs (train and val loss) of TinyDialogues using the typical iterative training approach for 20 epochs, for different global orders. From top to bottom: age order, reverse order, random order.

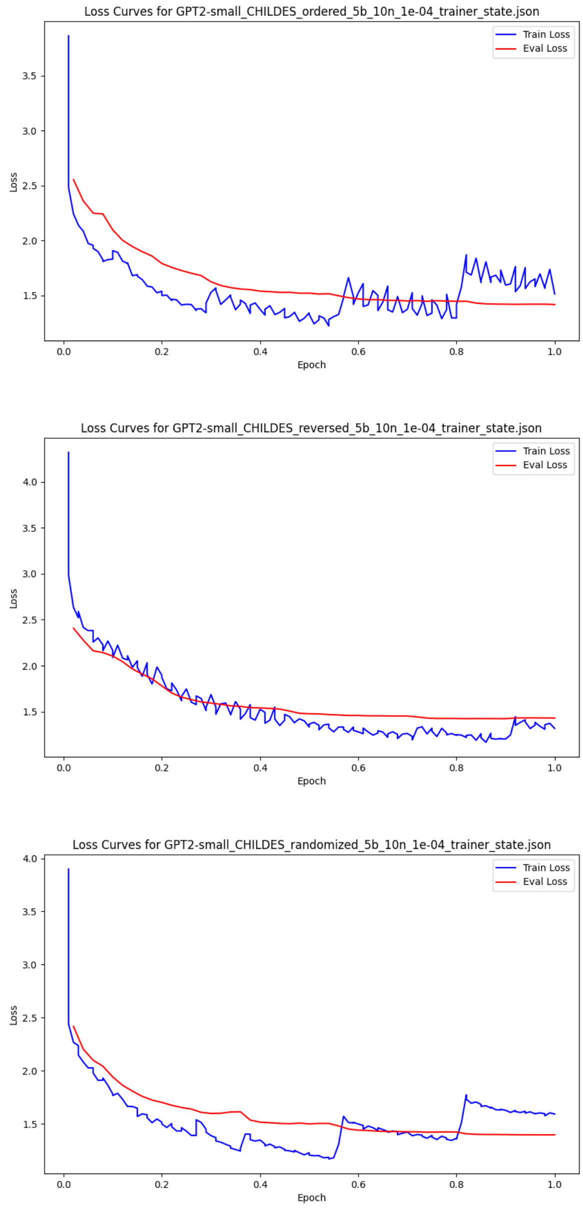

Figure 4: Convergence graphs (train and val loss) of CHILDES using the repeated buckets training approach with $b = 5, n = 10$, for different global orders. From top to bottom: age order, reverse order, random order.

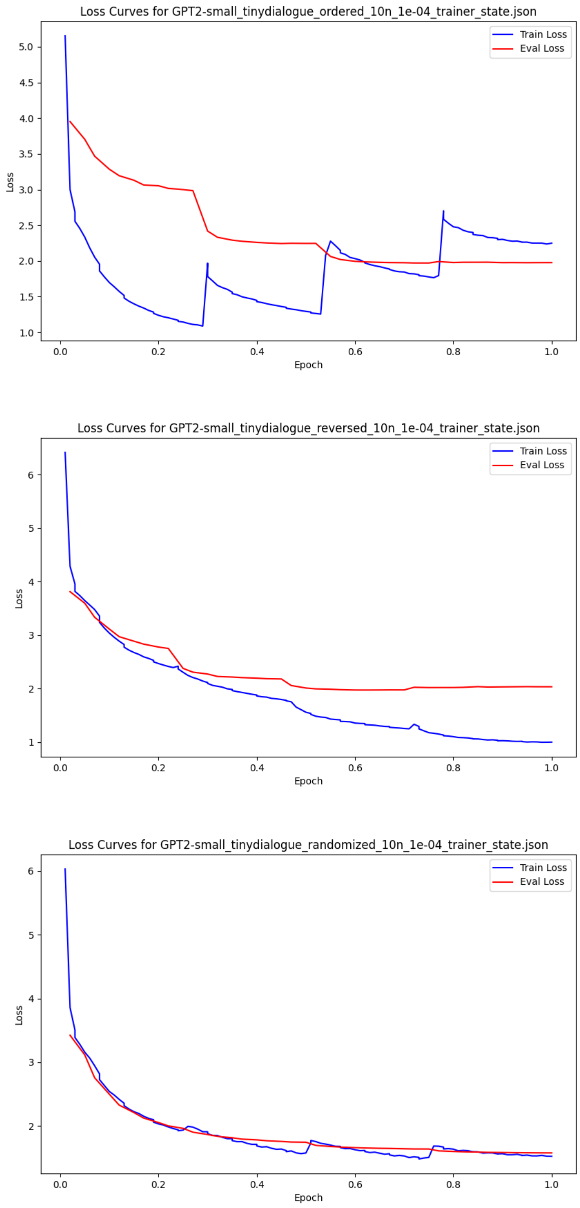

Figure 5: Convergence graphs (train and val loss) of TinyDialogues using the repeated buckets training approach with $n = 10$, for different global orders. From top to bottom: age order, reverse order, random order.

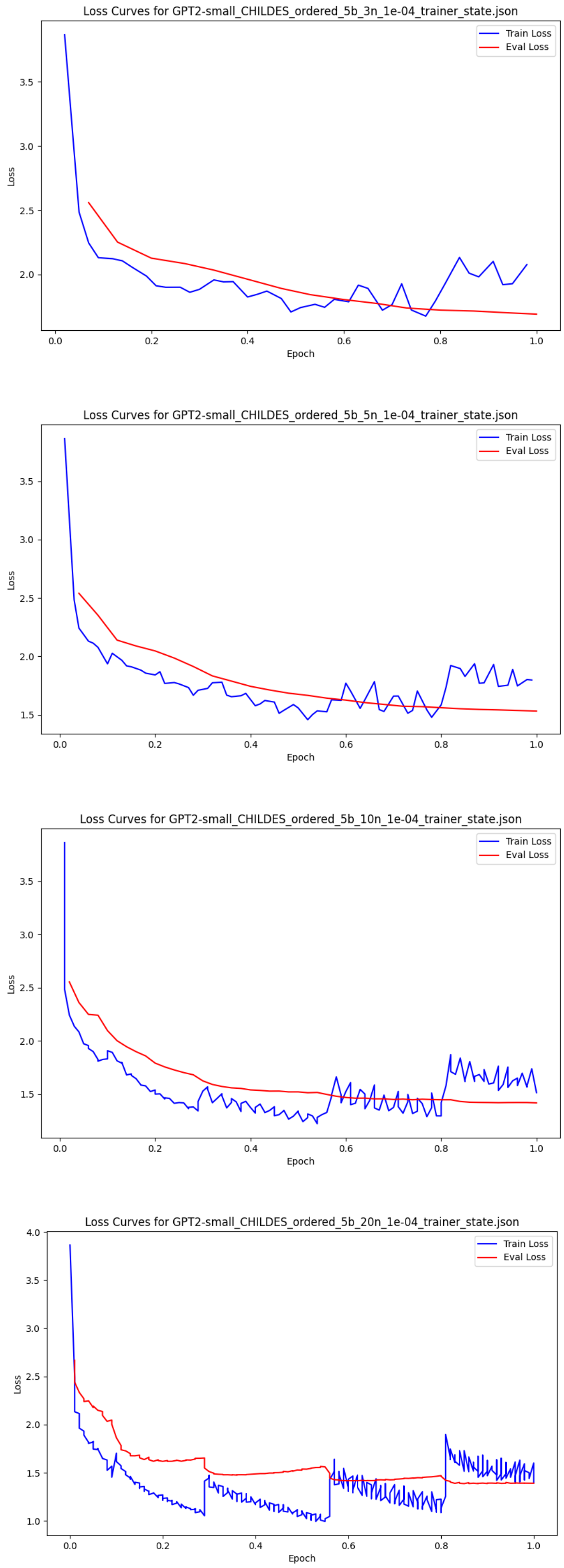

Figure 6: Convergence graphs (train and val loss) of CHILDES using the repeated buckets training approach with $b = 5$, for different values of n. From top to bottom: $n = 3, 5, 10, 20$.

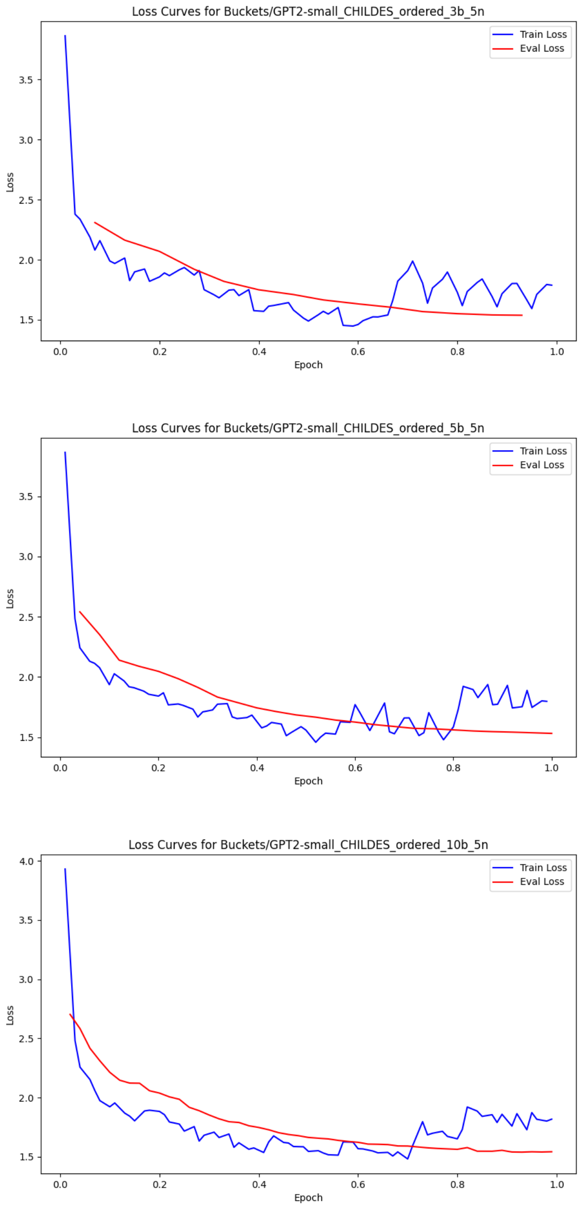

Figure 7: Convergence graphs (train and val loss) of CHILDES using the repeated buckets training approach with $n = 5$, for different values of b. From top to bottom: $b = 3, 5, 10$.

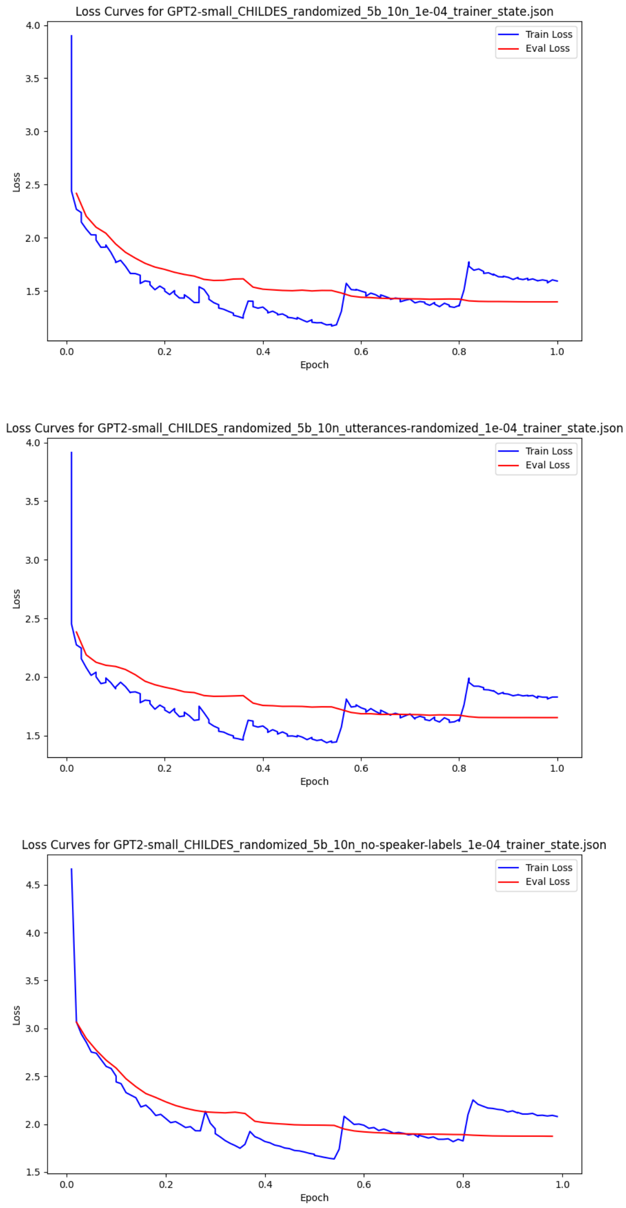

Figure 8: Convergence graphs (train and val loss) of CHILDES, looking at the effects of local interventions – shuffling utterances and removing speaker labels – using the repeated buckets approach with $b = 5, n = 10$. From top to bottom: original data (no changes), random shuffling of utterances, no speaker labels.