Child-directed speech facilitates production, not comprehension, in BabyLMs

Bastian Bunzeck and Sina Zarrieß
Computational Linguistics, Department of Linguistics
Bielefeld University, Germany
{bastian.bunzeck, sina.zarriess}@uni-bielefeld.de

Abstract

Recent studies suggest that child-directed speech is not conducive to language learning in BabyLMs. However, current evaluations focus predominantly on comprehension and not production, which is central to usage-based theories of language acquisition which argue how CDS facilitates early language use through constructional "frames" (frequent lexical patterns with open slots). We introduce a novel generation-based evaluation inspired by such theories in form of a **frame-completion task**, and compare Llama models trained with CDS, the BabyLM corpus, and web-crawl data (FineWeb-edu) on comprehension benchmarks and our novel framework. Our results reveal a clear dissociation between models' comprehension and production capabilities: while FineWeb-trained models excel at minimal pairs, CDS-trained models produce grammatical completions substantially earlier in training and concentrate probability mass on appropriate slot-fillers. These findings show that comprehension benchmarks underestimate what CDS affords to BabyLMs.[1]

1 Introduction

Child-directed speech (CDS) differs from regular speech, *inter alia*, through short utterances, exaggerated prosody and frequent repetition, properties that facilitate language acquisition by grabbing children's attention (Snow and Ferguson, 1977; Fernald, 1985; Soderstrom, 2007). Usage-based approaches to language acquisition argue that these distinctive distributional properties are significant for human learners, as they provide constructional frames which enable children to extract *productive* patterns for their own usage through item-based learning (Tomasello, 2000c; Diessel, 2013; Behrens, 2021) and frequency-driven mechanisms (Behrens, 2009; Diessel and Hilpert, 2016). This enables children to communicate their intentions effectively (Raz and Saxe, 2020), long before they have a full-fledged, adult-like grammatical system (Tomasello, 2000a).

Frame: I like to play with my _

Completion by model trained on data from:

CHILDES
▹ toys.

BABYLM
▹ own of the time.

FINEWEB-EDU
▹ be a few of the world.

FINEWEB-EDU-SHORT
▹ .

TINYDIALOGUES
▹ toys.

Figure 1: Lexical frame completions generated after 10% of training under greedy decoding, illustrating models' early production capabilities. More examples are presented in Appendix G.

Recently, work on LMs trained with acquisition-scale data (BabyLMs, Warstadt et al., 2023) has focused on the linguistic capabilities acquirable from CDS in neural learners, for example, by systematically manipulating the presence and amount of CDS in pretraining (e.g., Huebner et al., 2021; Padovani et al., 2025b; Bunzeck et al., 2025). Yet, the effect of CDS on language models remains contested. Although a few studies show clear positive effects of CDS on LMs' linguistic capabilities, like faster learning of syntactic specific phenomena (Huebner et al., 2021; Salhan et al., 2024), others show no advantage for CDS compared to wiki text (Padovani et al., 2025b) or a general inability to learn specific hierarchical rules (Yedetore et al., 2023). In practice, however, current BabyLM evaluation resources are limited to tests of implicit grammatical knowledge, e.g., measured with minimal pair datasets (Warstadt et al., 2020; Jumelet et al., 2026b). While informative about rule-based knowledge, these evaluations do not assess other aspects of linguistic knowledge that usage-based

[1] Models, prompts and other data can be found in this HuggingFace collection.

approaches are interested in (cf. Weissweiler et al., 2025), for example if models can produce acceptable and appropriate utterances. Further, while some work has mapped out the learning trajectories of comprehension-based benchmarks (cf. Choshen et al., 2022; Bunzeck and Zarrieß, 2024; Padovani et al., 2025b), almost nothing is known about how generation capabilities evolve during pretraining. Do models only produce word salad in their early checkpoints (cf. Figure 1), or do they generate utterances that are simple, short, and largely correct, as in children's speech? This distinction matters for usage-based and constructionist approaches, which predict that CDS facilitates the production of simple but appropriate early language use. Therefore, we hypothesize that the limited effects of CDS in LM training can be attributed to the restricted evaluation paradigm that is not tailored to capture conducive effects of CDS on **early** language use, which, however, is the central point in usage-based theories (Tomasello, 2003; Rowland et al., 2025).

To address this gap, we introduce our frame-completion task, a generation-based evaluation that measures productive capabilities. We pretrain models on a variety of more/less developmentally plausible datasets, prompt them to complete sentence fragments (corresponding to constructional frames), and analyze i) whether completions are grammatical, ii) how certain models are when completing frequent lexical frames, and iii) how generation capabilities develop during pretraining. Our findings support the assumptions made in usage-based linguistics: While models trained on complex corpora like FineWeb-edu[2] (Penedo et al., 2025) surpass models trained on CDS on minimal pair benchmarks, CDS-trained models produce more grammatical frame completions earlier in training. Furthermore, entropy analyses show that CDS models make more focused predictions, e.g., concentrating probability mass on semantically appropriate concrete nouns in argument positions. This pattern suggests CDS-trained models learn constructional templates that scaffold productive combination and the correct completion of lexical frames with appropriate lexical items early in pretraining, whereas models trained on web text attempt to generate overly complex syntactic structures. This divergence is hidden when *only* using minimal pairs to measure the linguistic knowledge of BabyLMs.

[2]We designate datasets in small caps (e.g., FineWeb-edu) and models trained on them in monospace (`FineWeb-edu`).

2 Related work

Usage-based language acquisition A central tenet of usage-based language acquisition is that linguistic knowledge is built up through *use* in an active learning process (Tomasello, 2000b; Raz and Saxe, 2020; Rowland et al., 2025). Using language means achieving communicative goals through its production (Diessel, 2017). Children do so long before their language system is fully abstract. Production then is not merely a reflection of acquired knowledge, but a driver of it. In the preverbal stage, children use grunts, babbling and pointing to achieve communicative goals (McCune, 2008; McGillion et al., 2017), which already improves motor skills necessary for production, before moving on to semantically motivated isolated words (e.g., *there*, *mommy*) and holistic phrases (e.g., *get-it*, *all-gone*). Early multiword utterances typically revolve around a 'pivot word' (a fixed anchor like *More* _ or *Want* _) with an open slot (cf. Braine and Bowerman, 1976; Lieven et al., 1997; Tomasello, 2000c) and serve particular speech-act functions (Tomasello, 1992, 2003). These utterances are reflected in the input: The vast majority of CDS utterances combine a highly frequent *frame* for the utterance with an open slot. For example, copular clauses are typically introduced by a pronoun and an auxiliary followed by a slot for nominals (e.g., *That's*/*It's* + ENTITY). Such frames lay the ground for children's early utterances (cf. Section 3).

CDS in BabyLMs Recent BabyLM studies have produced contradictory findings on the effect of purely CDS-based pretraining on linguistic knowledge. On the positive side, CDS has been shown to improve performance on various benchmarks (Zorro, BLiMP, cloze tests) over traditional pretraining data such as Wikipedia in masked and autoregressive LMs (Huebner et al., 2021; Qin et al., 2024; Feng et al., 2024), across languages (Salhan et al., 2024), as pretraining for further fine-tuning (Mueller and Linzen, 2023), for in-context learning (Deshpande et al., 2023; Muckatira et al., 2024) and for completing child-caretaker dialogues in contrast to fine-tuned LLMs (Levandovsky et al., 2025). Mixed to negative results are found for capturing hierarchical generalizations in question formation (Yedetore et al., 2023), predicting word learning trajectories (Ficarra et al., 2025), but also for minimal pair benchmarks when comparing to book text (Yam and Paek, 2024), when comparing to construction distributions in written and spoken language, where

Child speech frames	Freq.	FineWeb-edu frames	Freq.
I like to play with my _	22	*It is one of the most _*	122
and then you put it in _	22	*It is interesting to note that _*	106
but I don't know how to _	21	*All you have to do is _*	92
when I grow up I want _	19	*The reason for this is that _*	91
I don't know how to get _	17	*This is due to the fact _*	91
but I don't know where the _	17	*Note that depending on the number _*	90
and then put it in the _	17	*It can also be used to _*	84
and then they went to the _	16	*This is one of the most _*	77
and how I would do it _	16	*It should also be noted that _*	75
can I have a bit of _	15	*It is a good idea to _*	72

Table 1: Six-word frames used for prompting, including frequency in the dataset. For full list of frames see Table 6.

written profiles outperform spoken ones on MP benchmarks (Bunzeck et al., 2025), when comparing to Wiki data across languages (English, French and German, Padovani et al., 2025b), to 5-gram models (Vazquez Martinez et al., 2023), or when training on dialogue (Padovani et al., 2025a). These contradictions likely stem from differences in evaluation method and comparison baselines, but also from a shared reliance on comprehension-based tasks. This is partly architectural: Many BabyLMs are masked language models unsuitable for text generation, and even autoregressive ones lack the reinforcement-tuning that makes large models fluent generators.

Evaluating production Production-based evaluations have so far received little attention. Pannitto and Herbelot (2020) take an explicit usage-based stance and train LSTMs (not Transformer LMs) on 3M words of CHILDES data, OpenSubtitles, and Simple English Wikipedia. Across several checkpoints, they generate text and extract specific subtrees, from which they find that CHILDES-trained models approximate their input best, which they attribute to the repetitiousness of CDS. Nikolaus and Fourtassi (2021) show that for a language-and-vision task, production-based learning through corrective feedback improves performance over perception-only learning (note, however, that the evaluation method stays the same here). Using a masked language model trained on the BabyLM corpus (CDS + other data sources), Rozner et al. (2025b) show that such small models have high affinity between open slots and their appropriate fillers, even for abstract constructions such as *let-alone* and *much-less*. Lee and Berg-Kirkpatrick (2025) train autoregressive models on synthetic child stories and adult-oriented texts, and measure completion quality through traditional readability measures (e.g., Flesch-Kincaid score) and LLM-as-a-judge scoring. They find that models trained on supposedly more "readable" text generate coherent completions later, not earlier, than models trained on less readable data. In fact, learnability comes from less n-gram diversity, not readability measures. Levandovsky et al. (2025) train autoregressive models on dialogue from CHILDES and use an LLM-as-a-judge approach to measure completion coherence. Here, a model completely trained on CHILDES outperformed larger pretrained models that were only fine-tuned on CHILDES data, and a robot powered by this model was rated as more child-like by human test subjects. While these generation-based studies show mixed results for training on simple language or CDS, they respond to an emerging call for more production-based evaluation: In particular, Weissweiler et al. (2025) argue that evaluations should use natural stimuli and focus more on partially-filled schemas and constrained slots, exactly because current probing methods like MPs are not sufficient for constructionist approaches to language. Our study contributes to this line of work by introducing generation-based measures and asking if they converge on similar conclusions to comprehension-based measures like minimal pairs.

3 Lexical frames anchor production

Lexical frames in usage-based theory Usage-based linguistics generally considers frequent lexical patterns to be the principal input and output of early language acquisition. The first corpus-driven characterization of these lexical patterns was provided by Cameron-Faulkner et al. (2003), who investigated what they call "lexical frames", highly frequent utterance-initial lexical patterns that in-

troduce different types of elements for which they provide open slots (for example *What is _?* or *There _ go/es*). They found that over 65% of the utterances children hear are introduced by such frames, which they identified based on a study-specific frequency criterion (occurring at least 4 times per recording, used by at least half of the recorded mothers). Moreover, they showed that children's use of item-specific patterns is only loosely correlated with frames in the input, constrained to constructions simple enough for children, and often adapted, as in cases of deictic substitution between first- and second-person pronouns in dialogue. This shows that, while children learn from such patterns in an item-based fashion, they not only regurgitate the input but derive their own formulaic patterns. These results hold across languages and samples (cf. Stoll et al., 2009; Arnon, 2016; Bunzeck and Diessel, 2025).

Frame set for evaluation Given the prevalence of such frames in learners' input and output, we identify them as the ideal source of prompts for our generation evaluation. For that, we extract the most frequent utterance beginnings of four corpora: Child speech from CHILDES (MacWhinney, 2000) (which we do *not* use for pretraining), CDS from Childes (which we use as pretraining data); one subset of FineWeb-edu (Penedo et al., 2024) for target frames, and another subset (FineWeb-edu) that we use for pretraining. We assign frame status to n sentence-initial word sequences that occur at least 10 times in sentences with a length $> n$. All datasets feature 10M tokens, except the child speech from Childes, for which only 7M words are available. We then subtract the frames in the pretraining data from the set of frames in the held-out data. This means that there is *no* overlap in frames between training and evaluation data, and any frame that we use for generation is infrequent in the training data. This ensures that we test for frame completions on data that is at most infrequent in the training data, rather than on frames that models may simply memorize. Furthermore, children's own frames diverge from CDS frames, as discussed above and by Cameron-Faulkner et al. (2003). This leaves 11,000 candidate frames for the Childes child speech and 8,000 candidate frames for the held-out FineWeb-edu dataset.

To keep our prompting dataset concise, we select the top 10 most frequent 6-word, 5-word, 4-word, 3-word, and 2-word beginnings from both sets, manually discarding formulaic or boilerplate frames (e.g., Wikipedia boilerplate from page footers, nursery rhymes, or counting/spelling exercises). This results in 100 prompts (50 from Childes and 50 from FineWeb-edu). Table 1 displays the 6-word frames (full data in Table 6, Appendix C). For both data sources, the general frame structure is fairly similar: words are generally short, and the subject position is mostly filled by personal or demonstrative pronouns. Yet, there are also differences: The Childes frames include word order typical of questions, whereas the FineWeb-edu frames only introduce propositional sentences, frequently featuring a form of the copula verb *be*.

4 Pretraining

Data We pretrain our LMs on five different corpora of 10M lexical words (approximating the lexical input a child has received between 2 and 5 years of age, as per Warstadt et al., 2023 and Gilkerson et al., 2017): (i) the regular BabyLM corpus (Hu et al., 2024), which aims to represent the whole breadth of child-available input, (ii) child-directed speech (*not* child speech) from Childes, which maximizes developmental plausibility by only including data attested in child-caretaker interactions, and (iii) FineWeb-edu, which contains web crawls filtered for educational value and closely resembles the data that standard LLMs are trained on, while maintaining a higher diversity than e.g., commonly-used Wikipedia dumps. As additional comparisons, we also train models on the length-restricted FineWeb-edu-short, featuring sentences with at most six words, and on synthetic child-directed speech from TinyDialogues (Feng et al., 2024). Notably, these corpora differ tremendously in their number of types, ranging from 25,104 in TinyDialogues to 341,476 in FineWeb-edu-short (cf. also Appendix A).

Models We train autoregressive BabyLMs as they can be used for probability- *and* generation-based evaluations. We use a Llama architecture modeled after `SmolLM2-135M` (Allal et al., 2025), which prioritize depth over width (Petty et al., 2024; Gupta et al., 2025) and have been shown to perform strongly across a variety of benchmarks when compared to LMs of similar size. We train for three epochs and save checkpoints after 1% and 10% of pretraining and after every completed epoch, resulting in five checkpoints per model. We name models after their training data (`CHILDES`,

BabyLM, FineWeb-edu, FineWeb-edu-short, and TinyDialogues) and make them available on HuggingFace.

5 Evaluation

5.1 Comprehension

As comprehension-based benchmarks, we use BLiMP (Warstadt et al., 2020), which covers a wide array of semanto-syntactic phenomena with synthetic minimal pairs; Zorro (Huebner et al., 2021), which is based on BLiMP but restricted in vocabulary to lexical forms found in CHILDES; and MultiBLiMP (Jumelet et al., 2026b), which contains minimal pairs targeting agreement phenomena that were derived from existing UD-treebanks.

5.2 Production/frame completion

Decoding To complete our lexical frames, we prompt all 25 models and generate at most 32 new tokens. We use nucleus sampling ($p = 0.9$) with a temperature of 0.8, following best practices for open-ended generation (Holtzman et al., 2020; Zarrieß et al., 2021) that balance output diversity with textual coherence. Given our large prompt set (100 prompts) and number of models, we generate three completions per prompt rather than doing extensive sampling per prompt to prioritize the variety of prompts over depth per individual prompt. Appendix E contains results for different sampling strategies and temperature settings. For the sake of evaluation, we consider only the first complete generated sentence (marked by the first occurrence of sentence-final punctuation, which every generated data point contained without exception).

Quality of generations To assess the quality of the generated text, we use an LLM-as-a-judge strategy, following Lee and Berg-Kirkpatrick (2025) and Levandovsky et al. (2025). In comparison to other tasks, LLMs are fairly reliable in acceptability ratings (Bavaresco et al., 2025). In line with best practices (Lee et al., 2025), we frame our evaluation task as a binary decision for acceptability (yes/no). Because no further context is provided for the generated sequences, acceptability here entails being free of syntactic or utterance-internal semantic errors. As the first study of this kind, we restrict our focus to this more simplistic notion of acceptability, although it is also conceivable that reliable judge LMs could grade the generated utterances on further rubrics such as meaningfulness or communicative effectiveness, or even use more complex scoring systems such as magnitude estimation. We use qwen3-4b-2507 (Yang et al., 2025) as our judge model, as it provided consistent annotations that strongly align with expert judgements on a held-out test set and outperformed comparable models by a wide margin (cf. Appendix B for a comprehensive evaluation). We use this model to annotate all generated sequences with acceptability labels. Further, we calculate sequence length and TTR for all sequences as measures of lexical diversity.

Slot-wise productivity We annotate all 100 lexical frames with their "canonical" next element (NP, VP, clause, or unclear). To further analyze the development of our models' generative capabilities, we calculate two measures for next-token prediction across our frames: Shannon Entropy H and maximum probability P_{max}.

H quantifies the average uncertainty or "information content" in the model's predicted probability distribution. We calculate H as follows, with V being the vocabulary size and p_i being the probability of the i-th token: $H(P) = -\sum_{i=1}^{V} p_i \log_2(p_i)$.

The maximum probability P_{max} is an indicator of whether the model has a clear preference for the next token, calculated as the maximum value in the Softmax output: $P_{max} = \max_{i \in V}\{p_i\}$.

6 Results

Table 2 displays evaluation scores for the final models on our novel (generation-based) frame-completion task and established (comprehension-based) minimal pair evaluations. For the production-based measures, we bootstrap confidence intervals by drawing 10,000 resamples of the binary judgments and taking the 2.5th/97.5th percentiles of the resulting means (Efron and Tibshirani, 1993). Wilson 95% CIs (Wilson, 1927) were calculated as a sanity check and yielded near-identical values.

Production-based evaluation with lexical frames For text generation prompted with frames from Childes child speech, the CHILDES model trained on child-directed speech is the clear winner, producing 92.8% acceptable generations (Table 2). The synthetic TinyDialogues model also scores above 90%, and confidence intervals overlap considerably between these models. The BabyLM model, which also contains CDS from Childes, performs slightly worse at over 80%, while both web-trained models yield the lowest acceptability rates (44.1%

	Acceptable generations (CIs)		MP benchmarks		
Model	CHILDES	FineWeb-edu	BLiMP	ZORRO	MultiBLiMP
CHILDES	**92.8%** [88.8, 96.7]	44.1% [36.2, 52.0]	63.7%	73.8%	58.6%
BabyLM	80.9% [74.3, 86.8]	37.5% [29.6, 45.4]	67.4%	**81.0%**	82.6%
FineWeb-edu	44.1% [36.2, 52.0]	21.1% [14.5, 27.6]	**68.0%**	76.6%	**91.0%**
FineWeb-edu-short	59.9% [52.0, 67.8]	38.2% [30.3, 46.1]	61.9%	68.7%	52.9%
TinyDialogues	91.4% [86.8, 95.4]	**61.8%** [53.9, 69.1]	61.3%	68.8%	57.4%

Table 2: i) Proportions of acceptable frame completions (generated with a temperature of 0.8 and nucleus sampling with $p = 0.9$) with bootstrapped confidence intervals, and ii) accuracies on MP benchmarks for our Llama models trained on CHILDES, BABYLM, FINEWEB-EDU/-SHORT, and TINYDIALOGUES after three epochs.

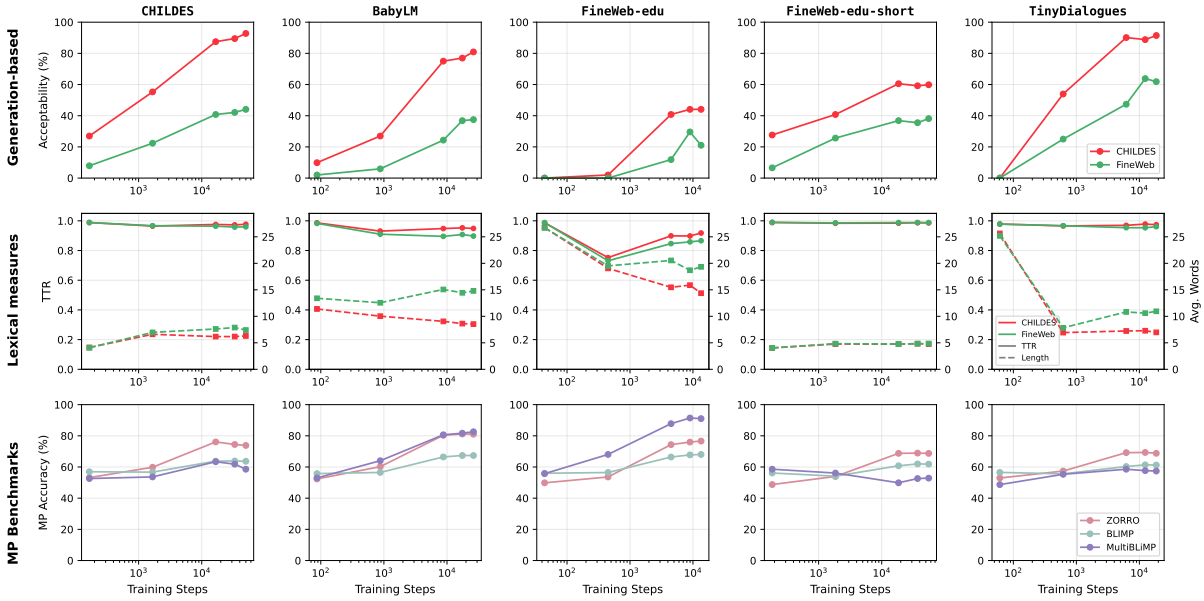

Figure 2: Development of acceptability of generated text, lexical measures, and MP benchmarks.

for FineWeb-edu, 59.9% for FineWeb-edu-short). This pattern might not seem overly surprising, given that models containing spoken data perform well on spoken-style frames. However, results on FINEWEB-EDU prompts show a similar pattern: The TinyDialogues model performs best (61.8%) and CHILDES also fares moderately well (44.1%). In contrast, the FineWeb-edu-short, with a maximum sentence length of 6, achieves only 38.2% (although confidence intervals overlap with CHILDES results). The regular FineWeb-edu model performs worst by a clear margin, at 21.1% (although pretraining data and prompts are drawn from the same underlying dataset). We further report pairwise statistical comparisons in Appendix D and provide a qualitative analysis of two three-word frames in Appendix G.

Comprehension-based evaluation Performance on the MP benchmarks follows a drastically different pattern (Table 2). While all models score above chance across all MP benchmarks, the FineWeb-edu model, which generates almost no acceptable text, performs best on BLiMP and MultiBLiMP. The BabyLM model achieves the highest score on Zorro. The best-generating CHILDES, on the other hand, only performs reasonably well on Zorro, but poorly on MultiBLiMP. Interestingly, the other strong generation model, TinyDialogues, also dramatically underperforms on all MP benchmarks.

Developmental trajectories Figure 2 displays the developmental trajectories for the evaluation measures discussed above. Generally speaking, frame-completion success on CHILDES and FINEWEB-EDU prompts improves in tandem across all models. Apart from that, the difference between CDS-trained models and models trained on traditional LM training data is striking. Both natural (CHILDES) and synthetic (TinyDialogues) models exhibit steep improvement curves between 1-

10% of pretraining. In contrast, the BabyLM and FineWeb-edu-short models improve markedly later. The regular FineWeb-edu is the most extreme outlier, as it only starts to generate acceptable utterances after one full epoch of pretraining.

This outlier status is also confirmed by the lexical measures (middle row of Figure 2). Generally, type-token ratio stays close to 1, indicating that models do not generate repeated words, except for FineWeb-edu, which tends to generate repetitive output early in pretraining. The length of generated sequences is more informative: It is low and remains low for CHILDES and FineWeb-edu-short, and somewhat longer for BabyLM. For FineWeb-edu and TinyDialogues, the picture differs substantially: both start out with very long sequences (over 25 words), which then decrease: FineWeb-edu stabilizes between 15-20 words, while TinyDialogues decreases to 5-10 words, comparable to CHILDES (cf. also examples in Figure 1).

The trajectories of the MP benchmarks are more uniform and start out at a higher level of performance (i.e., due to the nature of the MP scoring task). The largest improvements happen between 10-100% of pretraining, with the BabyLM and FineWeb-edu models improving the most. Improvements across all three benchmarks happen in tandem; only the FineWeb-edu-short remains an outlier with surprising performance decreases on MultiBLiMP.

Slot-wise entropy and maximum probability Figure 3 shows the development of Shannon entropy H and $P_{\max}$ for our 100 prompts, separated by the canonical slot fillers. While H generally decreases across training, $P_{\max}$ increases. For the CHILDES and TinyDialogues models, the final checkpoints see a slight increase in entropy without a decrease in $P_{\max}$. The highest entropy (≈ 7) is maintained by the FineWeb-edu model, followed by CHILDES and BabyLM (≈ 6), while FineWeb-edu-short and TinyDialogues reach the lowest scores (≈ 5). Regarding fillers, all models show the highest entropy for NPs. Lower entropies differ between VP (for CHILDES and TinyDialogues) and clause (for BabyLM, FineWeb-edu, and FineWeb-edu-short). Similar trends are visible for $P_{\max}$, where FineWeb-edu reaches the lowest average and FineWeb-edu-short the highest. Again, the VP category is the outlier for CHILDES and TinyDialogues, which assign highest $P_{\max}$ there.

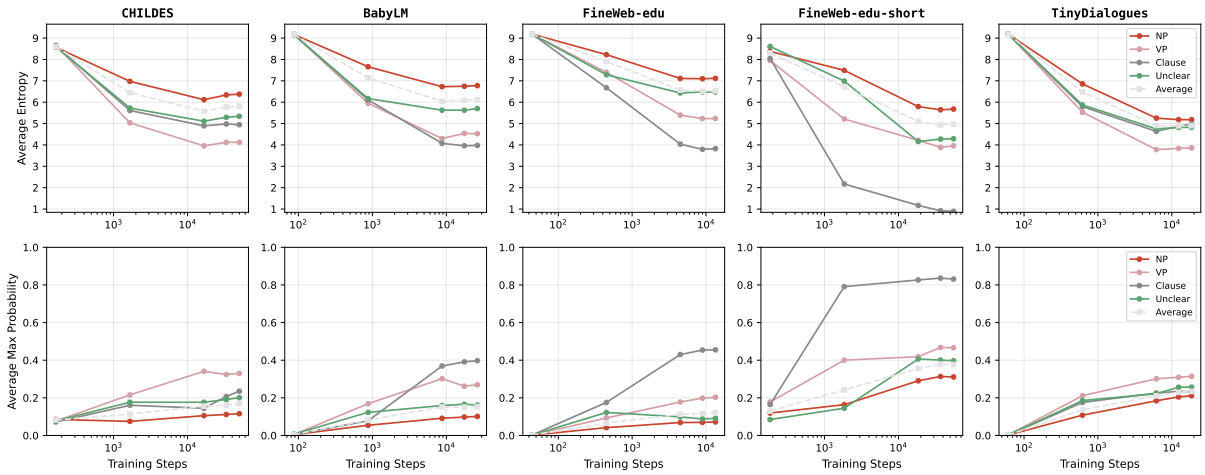

Figure 3: Development of slot-wise measures (entropy and max. probability), separated by canonical slot element.

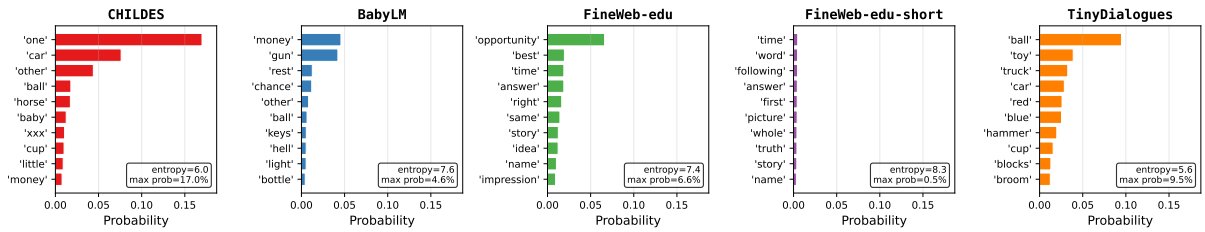

Figure 4: Next token predictions for *Give me the* at final checkpoint (3 epochs). Note that *xxx* is a CHILDES transcription convention for unintelligible speech.

Case study: *Give me the _* Figure 4 displays the top ten next-word predictions for the lexical frame *Give me the _* at the final checkpoint of all models (see Appendix F for intermediate checkpoints). The canonical element for this slot would be a noun phrase, for which all predicted tokens are valid candidates (mostly nouns, few adjectives). Regarding semantic properties, both CHILDES and TinyDialogues predictions focus on developmentally plausible words like *ball* and *car*. Interestingly, the CHILDES model also distinctively deviates from this pattern: the first and third most probable tokens are the discourse-deictic expressions *one* and *other* that are commonly used for tracking and distinguishing referents, which is a fundamental function of early referential communication (Tomasello, 2003). No other model predicts such words as particularly likely. In contrast, the FineWeb-edu models predict rather abstract terms (e.g., *time*, *answer*, *story*), and the BabyLM model's predictions appear to be heavily influenced by the unfiltered OpenSubtitles data in the corpus (*money* and *gun* are the most likely predictions).

Concerning slot-level metrics, there is (again) a clear distinction between CDS-trained and remaining models. Models trained on natural/synthetic spoken data are characterized by low entropy and high $P_{\max}$: TinyDialogues has the lowest entropy (5.6) and a $P_{\max}$ of 9.5%, whereas CHILDES entropy is somewhat higher (6.0), and $P_{\max}$ is 17%. In these models, probability distributions look most Zipfian (Zipf, 1935; Baroni, 2009), with probability mass being distributed across a small set of tokens with distinctive "winners". In contrast, the highest entropy is reported for FineWeb-edu-short, with many tokens being equally (un)likely at 0.5%. BabyLM and FineWeb-edu show similar high entropies (7.4–7.6) and low $P_{\max}$ (4.6%–6.6%). This means the web-trained models have less clear preferences for next tokens.

7 Discussion

In general, the results of our experiments confirm our hypothesis. The limited positive effects of CDS observed in prior BabyLM research can be plausibly attributed to restricted evaluation paradigms. When assessed on production instead of comprehension, CDS-trained models clearly outperform models trained on "typical" LM pretraining data. Four findings stand out: i) both CHILDES and TinyDialogues models achieve the highest acceptability rates (90–94%), both on CHILDES and on FINEWEB-EDU prompts, where the FineWeb-edu model trained on matching data actually underperforms (acceptability of only 21.1%). This suggests that mere exposure to grammatical data might be insufficient without the distributional tendencies that make slot-filling learnable. ii) CDS-trained models improve on frame completion earlier in training than web-trained models. iii) Entropy analyses reveal that CDS models develop Zipfian probability distributions with clear "winners", whereas web-trained models distribute probability mass across many tokens, reflecting uncertainty over appropriate tokens. iv) In the qualitative analysis of one frame, the predictions of the CHILDES model are semantically the most appropriate. While concrete nouns are also predicted by the TinyDialogues model, discourse-deictic expressions like *one* and *other*, which serve important referential functions in early communication (Tomasello, 2003), are only predicted by the CHILDES model.

Relation to prior work Our findings align with recent work questioning the relationship between evaluation and training data properties. Lee and Berg-Kirkpatrick (2025) show that reduced n-gram diversity in synthetic data predicts coherent generation;[3] similarly, our best-generating models are trained on the least lexically diverse, naturally-occurring data (cf. Appendix A). Levandovsky et al. (2025) corroborate this with CHILDES-trained models that outperform larger fine-tuned models on dialogue completion. From a learning-theoretic perspective, Kunstner and Bach (2025) show that long-tailed distributions challenge gradient-based learning. This suggests that CDS, with its smaller vocabulary and high-frequency frames, might provide more favorable distributions for pattern extraction. More generally, learnability might primarily depend on distributional structure rather than on content quality, as the educational, grammatically correct text of FineWeb-edu does not translate into early productive competence. Our findings also confirm previous results by Agarwal et al. (2025), who find no relationship between syntactic probes and BLiMP performance, further questioning whether different evaluation paradigms tap into the same underlying "knowledge".

Although their performance on benchmarks is roughly the same and individual differences might

[3] Unfortunately, the authors do not provide comprehension-based evaluation scores.

be due to the sample size, as indicated by the reported confidence intervals, the qualitative differences between CHILDES and TinyDialogues should be singled out once more. *Only* the CHILDES model predicts discourse-deictic expressions, which are central to acquisition but absent from TinyDialogues' predictions. This aligns surprisingly well with previous findings on synthetic text being linguistically much more uniform than natural language (Ju et al., 2025).

Implications for language acquisition Our results also connect to broader debates in usage-based acquisition research. The "Starting Big" approach (Arnon, 2021) argues that development proceeds from holistic chunks to analyzed components; our entropy analysis mirrors this trajectory, with CDS-trained models developing focused distributions over appropriate slot fillers, just as item-based learning predicts (Theakston and Lieven, 2017). Rowland (2007) find that children make more errors on questions not introduced by familiar lexical frames; the CHILDES model's success on CHILDES prompts can also be seen as confirming this pattern. Finally, the dissociation between production and comprehension we observe aligns with Tomasello (2000a), who argues against the assumption that children possess adult-like syntactic competence from the start. To quickly become effective language users, children do not need full knowledge of grammatical rules, but rather the means to produce appropriate words in appropriate contexts.

8 Conclusion

The present study demonstrated that comprehension-based evaluations alone underestimate what CDS offers to language learners. While our BabyLMs trained on complex, curated text perform well on established minimal pair benchmarks, our CDS-trained BabyLMs produce acceptable utterances earlier by completing lexical frames, and feature more focused predictions over slot fillers. This difference is crucial for usage-based linguistics, which predicts these patterns but still reports missing experimental evidence (Kempe et al., 2024), and for BabyLM research, where CDS has often been dismissed as unhelpful based on comprehension metrics. If one goal of BabyLM research is to approximate language acquisition, and acquisition proceeds through production, then generation-based evaluation should be as primary as comprehension-based evaluation, not merely supplementary.

Future work could expand upon our findings in several directions. As CDS is not static and constructions become more diverse (Bunzeck and Diessel, 2025) while redundancy decreases (Tal et al., 2024) across development, curriculum learning approaches that gradually increase input complexity could test further effects on production capabilities, especially in larger text-based models, as the web-text models (based on FineWeb-edu) showed the greatest deficiencies in early generation performance. Besides, our comparison focused on BabyLMs only. Since early checkpoints of larger models trained on trillions of tokens (such as OLMo2, Walsh et al., 2025) are now publicly available, analyses of such intermediate model variants could reveal whether the comprehension-production dissociation is specific to developmentally plausible small-data regimes like BabyLM, or also scales with model and data size. Finally, our binary acceptability classification discards all unacceptable utterances, and it would be highly interesting to see why these unacceptable utterances are wrong, if for the same reasons as children's ungrammatical utterances (Nikolaus et al., 2024), or if autoregressive LMs diverge in their early mistakes from those made by children.

Acknowledgments

We would like to thank Laurens Winkler for his help with training the base models and the anonymous CoNLL and CDL reviewers for their helpful comments and suggestions.

This research has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – CRC-1646, project number 512393437, project A02.

References

Ananth Agarwal, Jasper Jian, Christopher D Manning, and Shikhar Murty. 2025. Mechanisms vs. Outcomes: Probing for Syntax Fails to Explain Performance on Targeted Syntactic Evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33725–33745, Suzhou, China. Association for Computational Linguistics.

Liquid AI. 2025. LFM2 Technical Report. *arXiv preprint*.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček,

Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2025. SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model. *Preprint*, arXiv:2502.02737.

Inbal Arnon. 2016. The nature of CDS in Hebrew: Frequent frames in a morphologically rich language. In Ruth A. Berman, editor, *Trends in Language Acquisition Research*, volume 19, pages 201–224. John Benjamins Publishing Company, Amsterdam.

Inbal Arnon. 2021. The Starting Big approach to language learning. *Journal of Child Language*, 48(5):937–958.

Marco Baroni. 2009. Distributions in text. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics: An International Handbook*, pages 803–822. Mouton de Gruyter.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Heike Behrens. 2009. Usage-based and emergentist approaches to language acquisition. *Linguistics*, 47(2).

Heike Behrens. 2021. Constructivist Approaches to First Language Acquisition. *Journal of Child Language*, 48(5):959–983.

Martin D. S. Braine and Melissa Bowerman. 1976. Children's First Word Combinations. *Monographs of the Society for Research in Child Development*, 41(1).

Bastian Bunzeck and Holger Diessel. 2025. The richness of the stimulus: Constructional variation and development in child-directed speech. *First Language*, 45(2):152–176.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in German BabyLMs? In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 169–186, Vienna, Austria. Association for Computational Linguistics.

Bastian Bunzeck and Sina Zarrieß. 2024. Fifty shapes of BLiMP: Syntactic learning curves in language models are not uniform, but sometimes unruly. In *Proceedings of the 2024 CLASP Conference on Multimodality and Interaction in Language Learning*, pages 39–55, Gothenburg, Sweden. Association for Computational Linguistics.

Eduardo Calò, David M. Howcroft, Leo Leppänen, Saad Mahamood, Simon Mille, Patrícia Schmidtová, and Emiel Van Miltenburg. 2026. Justify Your Prompts! *Computational Linguistics*, pages 1–12.

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

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The Grammar-Learning Trajectories of Neural Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Vijeta Deshpande, Dan Pechi, Shree Thatte, Vladislav Lialin, and Anna Rumshisky. 2023. Honey, I Shrunk the Language: Language Model Behavior at Reduced Scale. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5298–5314, Toronto, Canada. Association for Computational Linguistics.

Holger Diessel. 2013. Construction Grammar and First Language Acquisition. In Thomas Hoffmann and Graeme Trousdale, editors, *The Oxford Handbook of Construction Grammar*. Oxford University Press, Oxford.

Holger Diessel. 2017. *Usage-Based Linguistics*. Oxford University Press.

Holger Diessel and Martin Hilpert. 2016. Frequency Effects in Grammar. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Bradley Efron and Robert John Tibshirani. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York.

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

Anne Fernald. 1985. Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, 8(2):181–195.

Filippo Ficarra, Ryan Cotterell, and Alex Warstadt. 2025. A Distributional Perspective on Word Learning in Neural Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11184–11207, Albuquerque, New Mexico. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul. 2017. Mapping the Early Language Environment Using All-Day Recordings and Automated Analysis. *American Journal of Speech-Language Pathology*, 26(2):248–265.

Akshat Gupta, Jay Yeung, Gopala Anumanchipalli, and Anna Ivanova. 2025. How Do LLMs Use Their Depth? *arXiv preprint*.

Joseph Henrich. 2024. WEIRD. In *Open Encyclopedia of Cognitive Science*, 1 edition. MIT Press.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

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

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Da Ju, Hagen Blix, and Adina Williams. 2025. Domain Regeneration: How well do LLMs match syntactic properties of text domains? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2367–2388, Vienna, Austria. Association for Computational Linguistics.

Jaap Jumelet, Abdellah Fourtassi, Akari Haga, Bastian Bunzeck, Bhargav Shandilya, Diana Galvan-Sosa, Faiz Ghifari Haznitrama, Francesca Padovani, Francois Meyer, Hai Hu, Julen Etxaniz, Laurent Prevot, Linyang He, María Grandury, Mila Marcheva, Negar Foroutan, Nikitas Theodoropoulos, Pouya Sadeghi, Siyuan Song, Suchir Salhan, Susana Zhou, Yurii Paniv, Ziyin Zhang, Arianna Bisazza, Alex Warstadt, and Leshem Choshen. 2026a. BabyBabelLM: A multilingual benchmark of developmentally plausible training data. In *Proceedings of the 19th Conference of the European Chapter of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 3297–3329, Rabat, Morocco. Association for Computational Linguistics.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026b. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Vera Kempe, Mitsuhiko Ota, and Sonja Schaeffler. 2024. Does child-directed speech facilitate language development in all domains? A study space analysis of the existing evidence. *Developmental Review*, 72:101121.

Frederik Kunstner and Francis Bach. 2025. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems*.

Ivan Lee and Taylor Berg-Kirkpatrick. 2025. Readability ≠ learnability: Rethinking the role of simplicity in training small language models. In *Second Conference on Language Modeling*.

Yukyung Lee, JoongHoon Kim, Jaehee Kim, Hyowon Cho, Jaewook Kang, Pilsung Kang, and Najoung Kim. 2025. CheckEval: A reliable LLM-as-a-Judge framework for evaluating text generation using checklists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15782–15809, Suzhou, China. Association for Computational Linguistics.

Enoch Levandovsky, Anna Manaseryan, and Casey Kennington. 2025. Learning to speak like a child: Reinforcing and evaluating a child-level generative language model. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 370–382, Avignon, France. Association for Computational Linguistics.

Elena V. M. Lieven, Julian M. Pine, and Gillian Baldwin. 1997. Lexically-based learning and early grammatical development. *Journal of Child Language*, 24(1):187–219.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, Baptiste Rozière, Baudouin De Monicault, Clémence Lanfranchi, Corentin Barreau, Cyprien Courtot, Daniele Grattarola, Darius Dabert, Diego de las Casas, Elliot Chane-Sane, Faruk Ahmed, Gabrielle Berrada, Gaëtan Ecrepont, Gauthier Guinet, Georgii Novikov, Guillaume Kunsch, Guillaume Lample, Guillaume Martin, Gunshi Gupta, Jan Ludziejewski, Jason Rute, Joachim Studnia, Jonas Amar, Joséphine Delas, Josselin Somerville Roberts, Karmesh Yadav, Khyathi Chandu, Kush Jain, Laurence Aitchison, Laurent Fainsin, Léonard Blier, Lingxiao Zhao, Louis Martin, Lucile Saulnier, Luyu Gao, Maarten Buyl, Margaret Jennings, Marie Pellat, Mark Prins, Mathieu Poirée, Mathilde Guillaumin, Matthieu Dinot, Matthieu Futeral, Maxime Darrin, Maximilian Augustin, Mia Chiquier, Michel Schimpf, Nathan Grinsztajn, Neha Gupta, Nikhil Raghuraman, Olivier Bousquet, Olivier Duchenne, Patricia Wang, Patrick von Platen, Paul Jacob, Paul Wambergue, Paula Kurylowicz, Pavankumar Reddy Muddireddy, Philomène Chagniot, Pierre Stock, Pravesh Agrawal, Quentin Torroba, Romain Sauvestre, Roman Soletskyi, Rupert Menneer, Sagar Vaze, Samuel Barry, Sanchit Gandhi, Siddhant Waghjale, Siddharth Gandhi, Soham Ghosh, Srijan Mishra, Sumukh Aithal, Szymon Antoniak, Teven Le Scao, Théo Cachet, Theo Simon Sorg, Thibaut Lavril, Thiziri Nait Saada, Thomas Chabal, Thomas Foubert, Thomas Robert, Thomas Wang, Tim Lawson, Tom Bewley, Tom Bewley, Tom Edwards, Umar Jamil, Umberto Tomasini, Valeriia Nemychnikova, Van Phung, Vincent Maladière, Virgile Richard, Wassim Bouaziz, Wen-Ding Li, William Marshall, Xinghui Li, Xinyu Yang, Yassine El Ouahidi, Yihan Wang, Yunhao Tang, and Zaccharie Ramzi. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3 edition. Lawrence Erlbaum Associates, Mahwah, NJ.

Lorraine McCune. 2008. *How Children Learn to Learn Language*. Oxford University Press.

Michelle McGillion, Jane S Herbert, Julian Pine, Marilyn Vihman, Rory dePaolis, Tamar Keren-Portnoy, and Danielle Matthews. 2017. What Paves the Way to Conventional Language? The Predictive Value of Babble, Pointing, and Socioeconomic Status. *Child Development*, 88(1):156–166.

Quinn McNemar. 1947. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157.

Sherin Muckatira, Vijeta Deshpande, Vladislav Lialin, and Anna Rumshisky. 2024. Emergent Abilities in Reduced-Scale Generative Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1242–1257, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Mueller and Tal Linzen. 2023. 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.

Mitja Nikolaus, Abhishek Agrawal, Petros Kaklamanis, Alex Warstadt, and Abdellah Fourtassi. 2024. Automatic annotation of grammaticality in child-caregiver conversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics,*

Language Resources and Evaluation (LREC-COLING 2024), pages 1832–1844, Torino, Italia. ELRA and ICCL.

Mitja Nikolaus and Abdellah Fourtassi. 2021. Modeling the Interaction Between Perception-Based and Production-Based Learning in Children's Early Acquisition of Semantic Knowledge. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 391–407, Online. Association for Computational Linguistics.

NVIDIA. 2025. Nemotron 3 Nano: Open, efficient mixture-of-experts hybrid Mamba-Transformer model for Agentic reasoning.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrylov, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpourlas, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. Gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint*.

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

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025b. Child-Directed Language Does Not Consistently Boost Syntax Learning in Language Models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19746–19767, Suzhou, China. Association for Computational Linguistics.

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: Evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849. Curran Associates, Inc.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language. *arXiv preprint*.

Jackson Petty, Sjoerd Steenkiste, Ishita Dasgupta, Fei Sha, Dan Garrette, and Tal Linzen. 2024. The Impact of Depth on Compositional Generalization in Transformer Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7239–7252, Mexico City, Mexico. Association for Computational Linguistics.

pleias. 2025. Baguettotron.

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

Gal Raz and Rebecca Saxe. 2020. Learning in Infancy Is Active, Endogenously Motivated, and Depends on the Prefrontal Cortices. *Annual Review of Developmental Psychology*, 2(1):247–268.

Caroline F. Rowland. 2007. Explaining errors in children's questions. *Cognition*, 104(1):106–134.

Caroline F. Rowland, Gert Westermann, Anna L. Theakston, Julian M. Pine, Padraic Monaghan, and Elena V.M. Lieven. 2025. Constructing language: A framework for explaining acquisition. *Trends in Cognitive Sciences*, page S1364661325001421.

Joshua Rozner, Leonie Weissweiler, Kyle Mahowald, and Cory Shain. 2025a. Constructions are Revealed in Word Distributions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2116–2138, Suzhou, China. Association for Computational Linguistics.

Joshua Rozner, Leonie Weissweiler, and Cory Shain. 2025b. BabyLM's First Constructions: Causal interventions provide a signal of learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2237–2249, Suzhou, China. Association for Computational Linguistics.

Suchir Salhan, Richard Diehl Martinez, Zébulon Goriely, and Paula Buttery. 2024. Less is more: Pre-training cross-lingual small-scale language models with cognitively-plausible curriculum learning strategies. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 174–188, Miami, FL, USA. Association for Computational Linguistics.

Hannah S. Sarvasy, Alan Rumsey, Josua Dahmen, John Onga, and Stephanie Yam. 2025. Child-directed speech in Ku Waru and Nungon (Papua New Guinea). *Language Development Research: An Open-Science Journal*, 5(3).

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and Statistical Modeling with Python. In *Python in Science Conference*, pages 92–96, Austin, Texas.

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

Melanie Soderstrom. 2007. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, 27(4):501–532.

Sabine Stoll, Kirsten Abbot-Smith, and Elena Lieven. 2009. Lexically Restricted Utterances in Russian, German, and English Child-Directed Speech. *Cognitive Science*, 33(1):75–103.

Shira Tal, Eitan Grossman, and Inbal Arnon. 2024. Infant-directed speech becomes less redundant as infants grow: Implications for language learning. *Cognition*, 249:105817.

Anna Theakston and Elena Lieven. 2017. Multiunit Sequences in First Language Acquisition. *Topics in Cognitive Science*, 9(3):588–603.

Michael Tomasello. 1992. *First Verbs: A Case Study of Early Grammatical Development*, 1 edition. Cambridge University Press.

Michael Tomasello. 2000a. Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Michael Tomasello. 2000b. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1-2):61–82.

Michael Tomasello. 2000c. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences*, 4(4).

Michael Tomasello. 2003. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press.

Hector Vazquez Martinez, Annika Lea Heuser, Charles Yang, and Jordan Kodner. 2023. Evaluating Neural Language Models as Cognitive Models of Language Acquisition. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 48–64, Singapore. Association for Computational Linguistics.

Evan Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–6, Singapore. Association for Computational Linguistics.

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

Leonie Weissweiler, Kyle Mahowald, and Adele E. Goldberg. 2025. Linguistic generalizations are not rules: Impacts on evaluation of LMs. In *Proceedings of the Second International Workshop on Construction Grammars and NLP*, pages 61–74, Düsseldorf, Germany. Association for Computational Linguistics.

Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158):209–212.

Hong Meng Yam and Nathan Paek. 2024. What should baby models read? Exploring sample-efficient data composition on model performance. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 284–291, Miami, FL, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. *arXiv preprint*.

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding Methods in Neural Language Generation: A Survey. *Information*, 12(9):355.

George K. Zipf. 1935. *The Psycho-Biology of Language*. Houghton Mifflin, Boston.

Limitations

Our study is accompanied by several limitations. First, we focus exclusively on English here. While CDS appears to be cross-linguistically widespread, its specific properties (e.g., frame frequencies, slot distributions) can be assumed to vary across languages. It is open to further inquiry whether our findings generalize to other languages that are, e.g., morphologically richer than English or allow more flexible word order patterns. Because CDS appears universal, it is present not only in WEIRD societies (Henrich, 2024), but also in communities such as the Kaluli of Papua New Guinea (Sarvasy et al., 2025). Cross-linguistic replication would strengthen claims about which properties of CDS matter, and remains a logical next step, especially with the release of datasets such as BabyBabelLM (Jumelet et al., 2026a).

Second, our LLM-as-a-judge approach, while validated against human annotations (Appendix B), could introduce potential biases. In line with the best practices outlined in Section 5, we focused on binary acceptability judgments, which might not capture finer-grained distinctions that methods like magnitude estimation could potentially help with. Also, as already outlined in Section 5.2, there are many more aspects of generation one could investigate, both more technical (such as fluency or readability) and more linguistic (such as effectiveness, appropriateness, the concrete nature of the unacceptable data, etc.). Future work could expand in this direction i) by investigating how child-like the productions really are, including possible linguistic mistakes in the data that may be more or less developmentally plausible, and ii) by looking into contextual appropriateness and more dialogue-based context instead of isolated sentences. In general, acceptability is a lower bound on what usage-based theory would call "appropriate" production. Conversely, this means that finer-grained measures (appropriateness, communicative adequacy) would likely widen and not narrow the gap between our examined models, since `FineWeb-edu`'s failure modes are also influenced by a register mismatch.

Third, our *as is* use of existing datasets limits causal isolation of the factors that drive performance in the frame-completion task. The complete disentanglement of aspects like register, lexical diversity, or frame-frequency requires further targeted ablations. However, existing patterns (length-controlled `FineWeb-edu-short` underperforms; diversity-controlled `TinyDialogues` lacks developmentally plausible discourse deictics) remain as non-trivial evidence against pure length or diversity explanations. Similarly, our experimental set-up does not control for shorter sub-frame frequencies or possible sub-frame n-gram leakage, an aspect that should be investigated in subsequent work.

Finally, we only evaluate autoregressive models in the current study due to their generative nature. While masked language models might not be straightforwardly usable with our approach, they are fairly common in BabyLM research and might show other comprehension-production differences stemming from their non-autoregressive pretraining goal. Here, it seems plausible that the methodology introduced by Rozner et al. (2025b,a) could be adapted to at least measure some of the aspects that we analyzed in free-form generation.

A Lexical distributions in the data

To give a more comprehensive overview of the different pretraining datasets, we compiled multiple measures of lexical diversity and overlap. Table 3

Dataset	n_{types}
CHILDES	40,200
BABYLM	109,631
FINEWEB-EDU	156,443
FINEWEB-EDU-SHORT	341,476
TINYDIALOGUES	25,104
Intersection	14,217

Table 3: Number of types in pretraining datasets (10M tokens each).

shows the absolute number of lexical types in our datasets. The web-crawled datasets FINEWEB-EDU and FINEWEB-EDU-SHORT feature a much higher lexical diversity with 156,443 and 341,476 respective types in the data. While the BABYLM data, which also contains text from the Simple English Wikipedia, Project Gutenberg, and OpenSubtitles next to CDS from the CHILDES corpora, is still quite diverse (109,631 types), the CDS corpora feature considerably fewer types. Our CHILDES dataset contains only 40,200 lexical types, and the synthetic TINYDIALOGUES dataset contains even fewer (25,104). Interestingly, this aligns with findings of synthetic data being less linguistically diverse than natural data (Ju et al., 2025) – the only synthetic dataset of our study is by far the least diverse. The intersection between all datasets amounts to 14,217 lexical types.

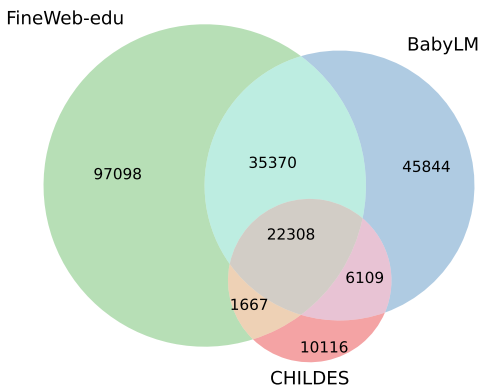

Figure 5: Lexical overlap between three pretraining corpora.

For more precise intersection statistics, Figure 5 shows the word-level overlap between our FINEWEB-EDU, CHILDES and BABYLM datasets in a Venn diagram, whereas Figure 6 shows overlap statistics for all corpora in an UpSet plot. The Venn diagram shows that the CHILDES dataset only contains 10,116 exclusive types when compared against FINEWEB-EDU and BABYLM, which both feature many more exclusive types and an overlap that is almost as large as the CHILDES dataset.

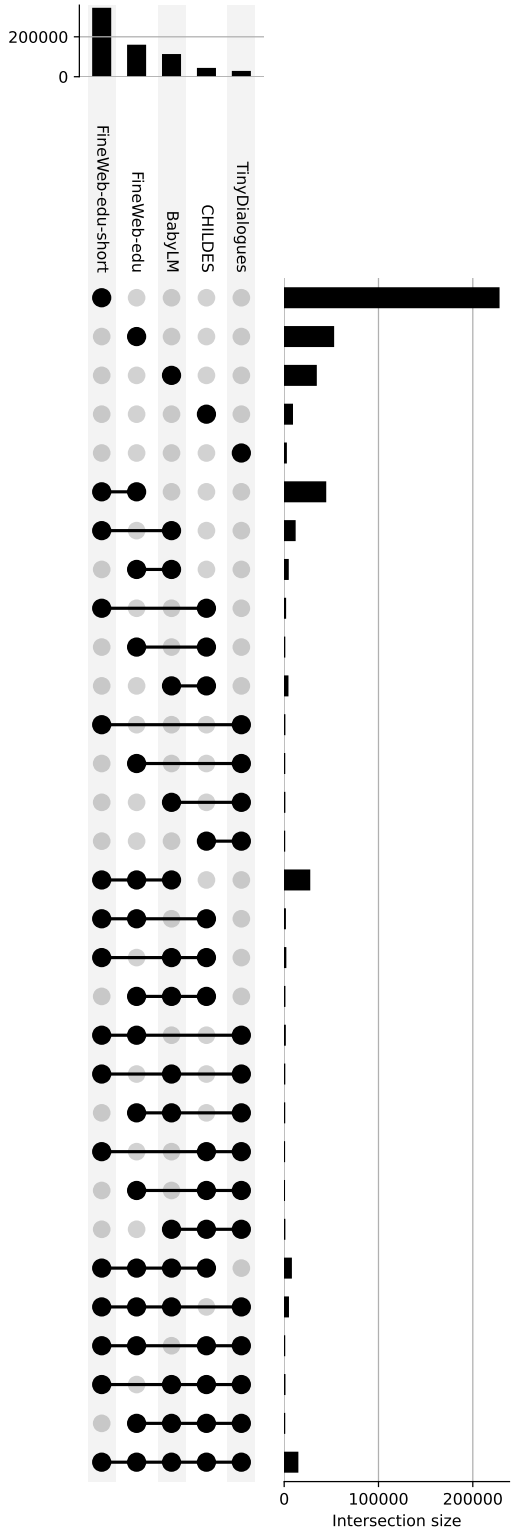

Figure 6: Lexical overlap between all pretraining corpora.

The UpsetPlot further visualizes overlap between

corpora. FineWeb-edu and FineWeb-edu-short share a quite large overlap with one another and also with the BabyLM data. Interestingly, there is no such tendency for our Childes data and the TinyDialogues dataset. This means that despite them both instantiating the same register – namely child-directed speech – there is little overlap. As such, it remains questionable if TinyDialogues is an accurate representation of CDS or not.

B LLM-as-a-judge

To determine an LLM that is suitable as a judge, we conducted two different runs of experiments, i) an open-ended qualitative examination and prompt optimization and ii) a systematic, quantitative study with human annotations.

Qualitative examination As a first step, we tested a variety of openly available, locally usable LMs via LM Studio. To do so, we took a small sample (10 utterances) from our generated sentences and tested different prompts with the following models: `qwen3-4b-2507` and `qwen3-4b-thinking-2507` (Yang et al., 2025), `gpt-oss-20b` (OpenAI et al., 2025), `nemotron-3-nano` (NVIDIA, 2025), `gemma-3-4b` (Gemma Team et al., 2025), `baguettotron` (pleias, 2025), `olmo-2-1124-7b` (Walsh et al., 2025), `ministral-3-3b` (Liu et al., 2026) and `lfm2.5-1.2b-instruct` (AI, 2025).

The first challenge is constructing a suitable prompt. As we request a binary judgement, it is important to precisely clarify what exactly should be judged. Here, we experimented with different formulations ("grammatically acceptable", "acceptable in spoken English"). Analyses of thinking traces and requests for clarifications and reasoning yielded a lack of context, a dispreference for short and informal utterances, as well as a lack of punctuation as reasons for rejection. Although thinking traces and post-hoc explanations cannot be taken as definitive evidence for the models' internal reasoning processes, they proved to be quite useful in prompt refinement. After several rounds of experimentation, we settled on the following prompt:

```
Is the following text grammatically acceptable and sound in English? Ignore punctuation for your judgement, please ONLY respond "yes" or "no": [Utterance to be evaluated]
```

Following this initial round of testing, we decided to exclude thinking models and comparatively large models, as the amount of generated tokens and general speed were unsuitable for the number of examples we evaluated. For the quantitative evaluation, we therefore focused on `qwen3-4b-2507`, `gemma-3-4b`, `ministral-3-3b` and `lfm2.5-1.2b-instruct`.

Quantitative evaluation To evaluate the narrowed selection of models, we manually annotated a random sample of 1000 completions with binary acceptability labels (yes/no). This resulted in a dataset with 646 unacceptable and 354 acceptable text strings. We tested the four target models with our best-performing prompt. Table 4 displays accuracy, precision, recall and F1 scores. `qwen3-4b-2507` is the clear winner with 95.9% accuracy and balanced precision/recall. In comparison, `gemma-3-4b` is decent but misses roughly 32% of acceptable sentences (lower recall). `ministral-3-3b` seems to be fairly conservative in its judgements, as it only predicts positive acceptability 39 times in total, but when it does it's almost always correct (97.4% precision). Finally, `lfm2.5-1.2b-instruct` overall shows mediocre performance.

As `qwen3-4b-2507` emerged as the clear winner, we further evaluated it for robustness across several runs. The results are shown in Table 5. With a majority vote, general statistics remain comparable to the results from Table 4 (accuracy = 95.6%, precision = 92.8%, recall = 94.9%, F1 = 93.9%). In general, 97.1% of the decisions show unanimous agreement across all 4 runs and only 2.9% of samples showing any disagreement. For these reasons, we settled for `qwen3-4b-2507` as a robust LLM-as-a-judge model which is "good enough" in the sense of Calò et al. (2026).

C Overview of lexical frames used for prompting

Table 6 lists all lexical frames that we use as prompts for our generation-based evaluation paradigm. We extract the top 10 most frequent 6-word, 5-word, 4-word, 3-word and 2-word frames that occur i) in child speech from CHILDES (but not in the CDS dataset Childes that we use for pretraining) and ii) in a 10M-word, randomly-sampled subset of FineWeb-edu (but not in the randomly-sampled FineWeb-edu dataset we use for pretraining).

An interesting difference between the frames from CHILDES and FineWeb-edu lies in their frequency distributions across different frame lengths.

Model	Accuracy	Precision	Recall	F1
`qwen3-4b-2507`	95.9%	93.8%	94.6%	94.2%
`gemma-3-4b`	85.9%	89.9%	67.8%	77.3%
`lfm2.5-1.2b-instruct`	70.9%	67.6%	34.2%	45.4%
`ministral-3-3b`	68.3%	97.4%	10.7%	19.3%

Table 4: Comparison of judge model performance.

Run	Accuracy
Original run	95.9%
Replication 1	95.9%
Replication 2	95.1%
Replication 3	95.1%

Table 5: `qwen3-4b-2507` accuracy across different runs.

For the 6-word frames, FineWeb-edu frames are much more frequent, occurring 70-120 times, whereas the CHILDES frames appear only 15-22 times. This pattern shifts considerably with shorter frames: the top 5-word CHILDES frame occurs 224 times, compared to only 141 occurrences for the top 5-word FineWeb-edu frame. For 3-word and 2-word frames, CHILDES frames are twice/thrice as frequent as their FineWeb-edu counterparts. This reflects the fact that long sequences are more prevalent in web data, whereas spoken utterances, especially child utterances, are simply not that long. Conversely, shorter utterance frames are more frequent in CHILDES data, consistent with the observation that lexical frames are highly frequent in spoken language.

D Pairwise comparison

To further test the robustness of our results, we provide pairwise comparisons for our generation-based evaluation. Pairwise comparisons use the exact binomial McNemar test (McNemar, 1947) on paired judgments, with Holm-Bonferroni correction (Holm, 1979). We use the `statsmodels` (Seabold and Perktold, 2010) implementation to carry out these tests. As Table 7 shows, most differences can be considered significant. Non-significant p-values cluster around models trained on similar data: for completions based on CHILDES frames, this concerns `CHILDES`, `BabyLM` and `TinyDialogues` as well as `FineWeb-edu` and `FineWeb-edu-short`, which further confirms register-based differences. For completions based on FineWeb-edu frames, non-significant pairwise comparisons are reported for `CHILDES`, `BabyLM` and `FineWeb-edu-short`, showing similarities based on the length of sentences in the input, although `TinyDialogues` is absent, which is somewhat surprising.

E Generation with different sampling strategies

Figure 7 compares the percentage of acceptable generations under different sampling paradigms and for both prompt sets. We test greedy decoding, top-k sampling ($k = 50$) and nucleus sampling (same as in the main results section, $p = 0.9$).

Greedy decoding does not dramatically deviate from other strategies and leads to competitive or best scores for most models, with very similar developmental trajectories across all sampling paradigms. Lower temperatures generally lead to more acceptable generations: a temperature of 0.4 frequently results in the highest acceptability rates, whereas a temperature of 1.2 produces the least acceptable outputs. Overall, nucleus sampling achieves the best overall scores. Importantly, no principal difference emerges between prompt sets, with Childes prompts consistently eliciting more acceptable generations than FineWeb-edu prompts, mirroring the patterns observed in our main results. These findings suggest that our results are robust and largely independent of the specific sampling strategies.

F Developmental trajectories in case study: *Give me the _*

Figure 8 displays the top ten next-word predictions for the lexical frame *Give me the _* across all five checkpoints of our models, which exhibit interesting developmental differences. The `CHILDES` model initially predicts punctuation with high likelihood, but then rapidly transitions to predicting fitting nouns at the next checkpoint (10% of pretraining). Its entropy trajectory evolves from a peaked distribution at first, through a flatter intermediate phase, to a well-formed Zipfian distribution. The `TinyDialogues` model follows a similar pattern, though without the early peak, ultimately also converging on a Zipfian distribution.

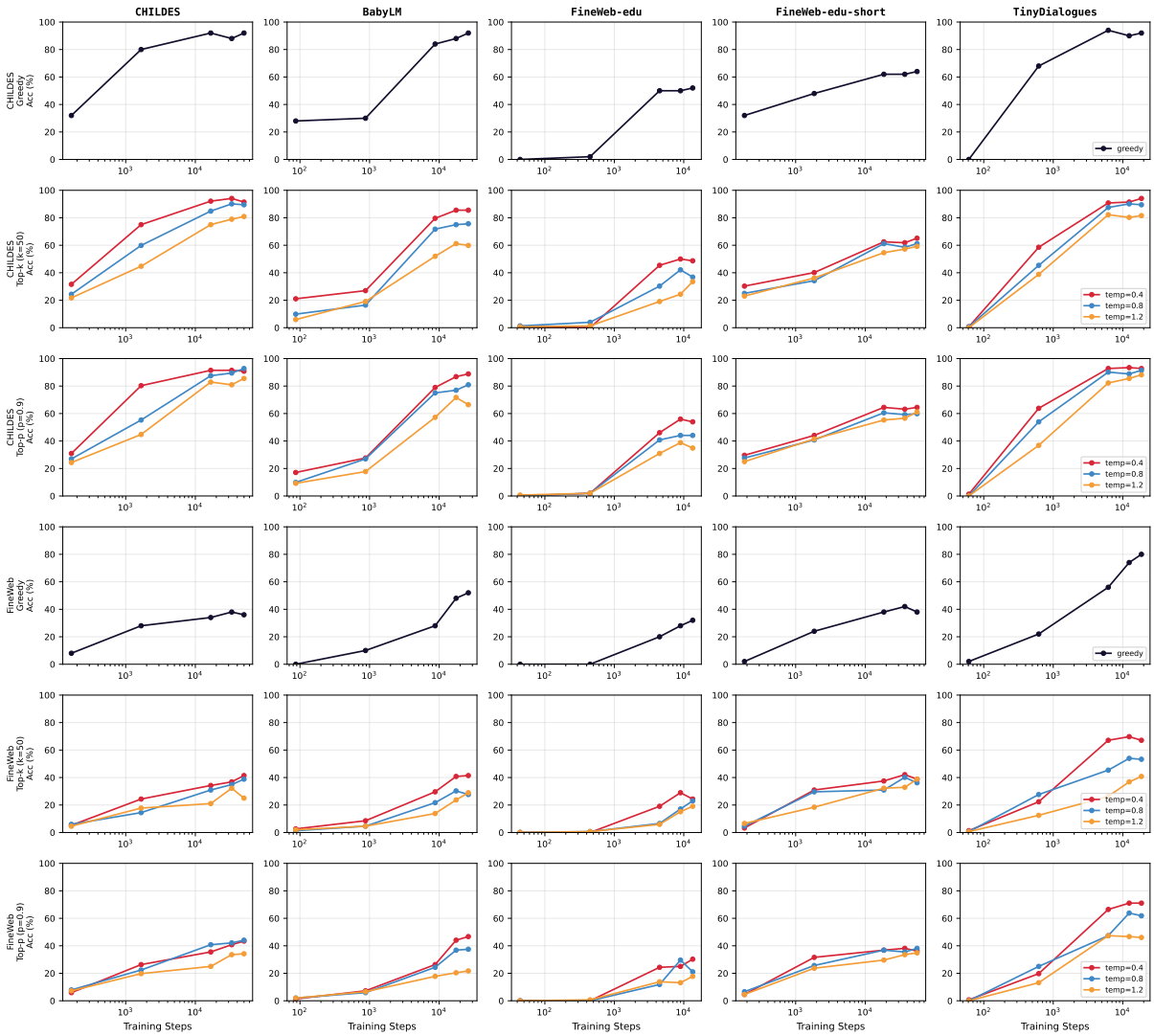

Figure 7: Acceptability statistics for different temperature settings and sampling strategies

In contrast, the `BabyLM` and `FineWeb-edu` models start with extremely high entropy and low maximum probability. As training progresses, one or two clear "winners" emerge in their predictions and remain relatively stable, but the overall distribution never becomes as cleanly Zipfian as for the CDS-trained models. Finally, the `FineWeb-edu-short` model is an outlier: It initially predicts punctuation with high probability, but these probabilities subsequently diminish and never consolidate again, resulting in many words receiving similarly low likelihoods throughout the remaining checkpoints.

G More examples

Tables 8 and 9 show more sample completions for two three-word frames: *look at the* _ from the English CHILDES corpora and *In short, the* _ from the FineWeb-edu dataset. Here, more differences between the developmental trajectories of the models become apparent. Across both lexical frames, patterns are strikingly similar. The `CHILDES` and `FineWeb-edu-short` models start with punctuation-only frame completions, which stabilize to short sentences, albeit only acceptable for the `CHILDES` data. In contrast, the `BabyLM` and `FineWeb-edu` models first generate longer strings of subword tokens after 0.01 epochs, which do not always represent lexical words. For `BabyLM`, this pattern then stabilizes to shorter, but ultimately grammatical sentences, at least after three epochs. The model trained on synthetic dialogues, `TinyDialogues`, exhibits a developmental pattern that is located somewhere between these extremes, starting out with extremely long strings at the first checkpoint (0.01 epochs), but then immediately providing short, acceptable strings after 0.1 epochs of training, which stay reasonably short.

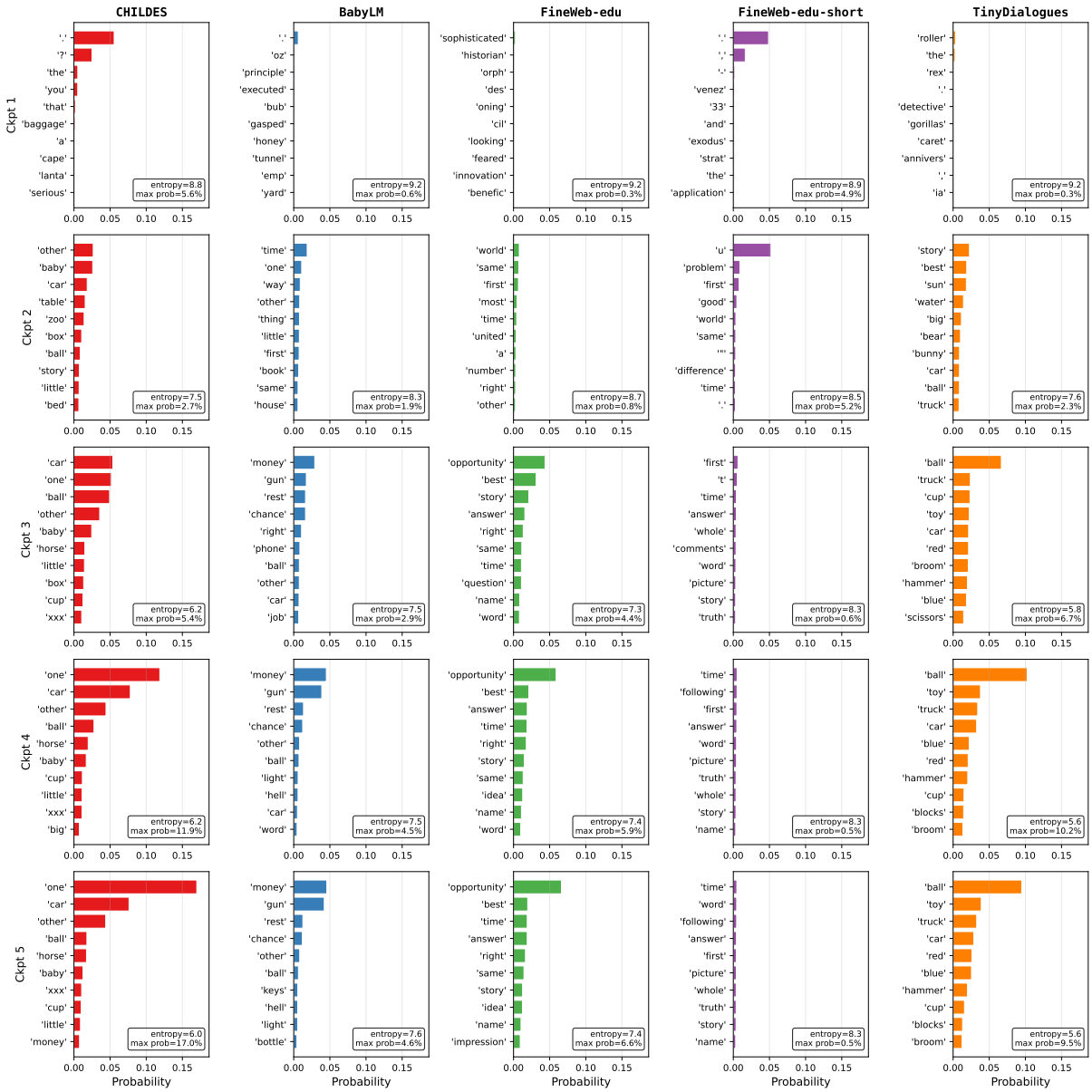

Figure 8: Next token predictions for *Give me the* at all checkpoints (0.01 epochs, 0.1 epochs, 1 epoch, 2 epochs, 3 epochs) of all five models.

n_{words}	Child speech frames	Freq.	FineWeb-edu frames	Freq.
6	*I like to play with my _*	22	*It is one of the most _*	122
6	*and then you put it in _*	22	*It is interesting to note that _*	106
6	*but I don't know how to _*	21	*All you have to do is _*	92
6	*when I grow up I want _*	19	*The reason for this is that _*	91
6	*I don't know how to get _*	17	*This is due to the fact _*	91
6	*but I don't know where the _*	17	*Note that depending on the number _*	90
6	*and then put it in the _*	17	*It can also be used to _*	84
6	*and then they went to the _*	16	*This is one of the most _*	77
6	*and how I would do it _*	16	*It should also be noted that _*	75
6	*can I have a bit of _*	15	*It is a good idea to _*	72
5	*and then there was a _*	224	*To learn more about the _*	141
5	*I don't know what it _*	76	*The bottom line is that _*	104
5	*I don't want you to _*	73	*It can be used to _*	102
5	*I wanna play with the _*	72	*At the same time the _*	98
5	*I don't know where it _*	54	*There are three types of _*	96
5	*I don't know what to _*	53	*For example, if you are _*	92
5	*I wanna go to the _*	52	*At the start of the _*	91
5	*what do you do with _*	49	*In the middle of the _*	89
5	*I want to play with _*	45	*This can be done by _*	88
5	*the boy and the dog _*	45	*What are the benefits of _*	84
4	*what is that funny _*	194	*However, there is a _*	143
4	*and this is a _*	172	*This allows you to _*	133
4	*how do you spell _*	161	*While there is no _*	118
4	*do you want a _*	151	*In recent years, the _*	114
4	*I think it's a _*	128	*That is why the _*	108
4	*it looks like a _*	122	*There are also some _*	104
4	*because I don't like _*	115	*In any case, the _*	102
4	*no I don't want _*	108	*In many cases, the _*	102
4	*put it on the _*	100	*Thank you for your _*	102
4	*what do you wanna _*	93	*In other words, a _*	98
3	*look at the _*	369	*I am a _*	180
3	*and that's the _*	302	*In short, the _*	176
3	*I like the _*	292	*Learn about the _*	159
3	*is that the _*	251	*The system is _*	141
3	*that is a _*	233	*He said the _*	140
3	*it is a _*	225	*Many people have _*	136
3	*a lot of _*	185	*Several of the _*	135
3	*and I was _*	181	*In conclusion, the _*	133
3	*that's a big _*	175	*He became a _*	132
3	*what is it _*	169	*Because of these _*	129
2	*to the _*	432	*The police _*	166
2	*a baby _*	342	*Using these _*	160
2	*not that _*	316	*Get your _*	156
2	*another one _*	236	*Eventually the _*	154
2	*that's your _*	235	*So to _*	142
2	*it a _*	233	*Accordingly, the _*	140
2	*where's your _*	224	*Otherwise, the _*	139
2	*take a _*	217	*This simple _*	135
2	*a red _*	196	*Consider a _*	135
2	*do that _*	192	*Since a _*	133

Table 6: Full list of lexical frames used for prompting.

		CHILDES frames				
		CHILDES	BabyLM	FineWeb-edu	FineWeb-edu-short	TinyDialogues
FW-edu frames	CHILDES	—	< 0.001	< 0.001	< 0.001	0.286
	BabyLM	0.625	—	< 0.001	< 0.001	0.002
	FineWeb-edu	< 0.001	< 0.001	—	0.002	< 0.001
	FineWeb-edu-short	0.625	1	< 0.001	—	< 0.001
	TinyDialogues	< 0.001	< 0.001	< 0.001	< 0.001	—

Table 7: p-values for pairwise comparisons. The upper right corner shows values for frames taken from CHILDES, the lower left corner shows values for FineWeb frames.

Epoch	CHILDES	BabyLM	FineWeb-edu	FW-edu-short	TinyDialogues
0.01	look at the.	look at thelwwwoz movie financial develop.	look at theicillinva healthy fully michel the beaut emblem combined spontaneous peculiaroping symbol journey the ppourage lim attractcy stand silicon tea gate ounkeley commit socioaga avoidingribeta	look at the.	look at the stopping professional leopards scaven national atmos pages trim mus lul unimp underwater massive spooked uneven gamble swoos the the spotting coconut bottomless tom parachutes pradogs disor recess admired pawn quarrel tasting
0.1	look at the one that's here.	look at the same one of the top.	look at the city of a first, it is not the u of the world.	look at the result.	look at the toy!
1	look at the little man.	look at the thing.	look at the other two things you can do.	look at the example below.	look at the flowers, sweetie.
2	look at the bear.	look at the, the big man!	look at the best way to prevent the problems that are important for your body.	look at the following.	look at the flowers.
3	look at the other page.	look at the two of us, and you'll be the one we've got.	look at the same number of species that can be used to build a single plant, but when you are out there, the researchers have discovered that the water is not a good	look at the following: 1.	look at the sun outside, it's so bright today!

Table 8: Additional sample frame completions for the lexical frame *look at the* _ from the English CHILDES corpora. Text generated with a temperature = 0.8 and nucleus sampling (p = 0.9), example randomly drawn from generated completions.

Epoch	CHILDES	BabyLM	FineWeb-edu	FW-edu-short	TinyDialogues
0.01	In short, the?	In short, thestay bub shore strongly gus italian disp music appliedhovenhoven conf software mann coffee the toadneyney fa bicy coloring mikeuesday elevator marsairs deeds du majesty magazheaded	In short, the psychologists afternoon ions 17 historianvolume lastsning resour relax poe exit thom6 retin pace).	In short, the,.	In short, the eelook whenever still the commitments maybebound twisting stays enslavescribbles national bun the unsha storybooks binsiclegobble seesaw blower bullseye checkers the national from whenever jenny ahhhnight timesaver
0.1	In short, the other one.	In short, the, a new was not for this.	In short, the best has been the american in the present of the state of the world and the story.	In short, the u.	In short, the garden!
1	In short, the first time.	In short, the old man, who was a man and his friend, and was at the time of the last one.	In short, the amount of energy in the soil is in the bottom of the body.	In short, the u.	In short, the wind was really strong, and he decided to bring his ball back to his home.
2	In short, the dark.	In short, the boys must be so bright that the boys might have to go down the river, and it would be good for the little girl to go through all the girls.	In short, the current study suggests that it may have been done with the best way to reduce the amount of time required by the patient.	In short, the u.	In short, the magic bottle is like a superhero.
3	In short, the green one?	In short, the whole thing is in there.	In short, the researchers are working to help improve the development of the information about the risk of cancer.	In short, the u.	In short, the sky is so big!

Table 9: Additional sample frame completions for the lexical frame *In short, the* _ from the FineWeb-edu dataset. Text generated with a temperature = 0.8 and nucleus sampling ($p = 0.9$), example randomly drawn from generated completions.