GPT-wee: How Small Can a Small Language Model Really Get?

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

Abstract

In this model report, we present an alternative approach to improving language models through scaling up their architectures and training data. In contrast, we train significantly smaller GPT-wee language models for the CMCL and CoNLL shared task: the BabyLM challenge. Drawing inspiration from usage-based linguistics, specifically focusing on language acquisition factors such as frequency, word length, and lexical frames, we also conduct tests employing curriculum learning techniques. Our findings demonstrate that even very small models can achieve considerable proficiency in standard evaluation tasks, performing as good as or even better than much larger baseline models, both on zero-shot evaluation and tasks that require further fine-tuning. Our naïve curriculum approach, however, does not show any straightforward improvements, except for certain, very specific tasks. Overall, the results remain inconclusive and suggest interaction effects between model architecture, data make-up and learning processes that warrant further inspection.

1 Introduction

In recent years, language model-based NLP has witnessed remarkable advancements, surpassing numerous benchmarks and continuously achieving new breakthroughs through increasingly larger models. However, such large language models come with certain difficulties. As their size expands, they demand substantial amounts of computing power and training data, while also retaining a certain degree of opaqueness and consuming immense amounts of energy (Bender et al., 2021). Besides, their overblown and complex architectures hinder interpretability, while commonly used training data mostly comes from non-naturalistic sources such as book corpora, Wikipedia crawls, and web pages. Addressing these concerns, the BabyLM challenge (Warstadt et al., 2023b) emerges as an experimental test bed for "smaller" or more optimized (and possibly more cognitively plausible) models. By drastically reducing the allowed amount of training data compared to state-of-the-art models, and by sourcing it from more varied domains, it forces language model engineers to come up with new solutions that are not (only) grounded in increasing parameters, training data, and computing power. We respond to this challenge by exploring language models with drastically reduced GPT-2 architectures (Radford et al., 2019) and the value of curriculum learning (Bengio et al., 2009; Hacohen and Weinshall, 2019) in training them, inspired by findings from usage-based linguistics on the nature of child-directed and child speech. Among the submissions to the BabyLM challenge, our GPT-Wee models stand out in the sense that we did not implement intricate and highly complex learning strategies, but rather examined how much simple architectures can be reduced in size while still providing considerable performance. Our models feature some of the lowest, if not *the lowest* number of parameters among the submissions.

Elman (1993) discusses how learning processes (e.g. language acquisition) are tied to cognitive maturation, and how during these processes, increasingly complex human neural networks are confronted with increasingly complex input. With respect to the concrete nature of this linguistic input, usage-based research has shown that its vocabulary is compact and mainly concerned with children's immediate surroundings (Saxton, 2017). It features a high amount of fragmentary utterances and frequent, utterance-initial lexical frames (Cameron-Faulkner et al., 2003). In turn, children's earliest utterances also revolve around lexically highly specific pivot schemas and item-based constructions (Braine and Bowerman, 1976; Tomasello, 2000; Diessel, 2013a), which only gradually expand to more complex utterances. Due to the linguistically

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 35–46
December 6-7, 2023 ©2023 Association for Computational Linguistics

more diverse nature of training data in machine learning – in the case of the present challenge it is, for example, composed of realistic input from CHILDES (MacWhinney, 2000) and other sources like Wikipedia dumps or Open Subtitles – the common approach of providing it to the training algorithm in random order does not mirror a developmentally plausible input trajectory. To better understand the value of these two factors, the growing intricacy of natural neural networks and the growing diversity of input, we (1) explore artificial neural networks of increasing/decreasing complexity, and (2) experiment with ordering the training data according to its complexity and (with regard to child-directed speech) prototypicality.

In sum, we find that a reduction of key parameters, e.g. the number of hidden layers and attention heads or the vocabulary size, does not immediately materialize in detrimental effects. Only when they are *drastically* reduced, the performance is affected more strongly. Moreover, we find that the curriculum approach does not always increase performance, but indeed shows effects on the training and evaluation losses that warrant further inspection. To the shared task, we submit our medium-sized model, as we observe the best size–performance trade-offs for these variants. We submit the curriculum variant (with a vocabulary size of 8k) of this member of our model family, which we call GPT-wee[1] in honor of their *wee* architectures.

2 Language models and developmental plausibility

Cognitive maturation in the form of an increasing number of neurons (*nodes*) and synapses (*connections)* in human neural networks accompanies developmental processes, and thus also language acquisition (Elman, 1993). The learning mechanisms of current language models do not mirror this development. Their architecture is defined before the training process, and then the nodes' and connections' weights and biases are randomly initialized and finally optimized, often based on randomly ordered input examples and influenced by the choice of specific loss functions. Interestingly, alternative approaches to ANNs, like dynamically growing networks or weights with gradient values, which were proposed during the 1980s and 1990s (for example in Elman et al., 1996, 73), never achieved widespread adoption in NLP (although they exist, with examples like NEAT (Stanley and Miikkulainen, 2002) having been shown to be useful for a variety of tasks). The best proxy for investigating the effects of neuronal growth are smaller models like BabyBERTa (Huebner et al., 2021) or TinyStories (Eldan and Li, 2023). They show that for small data settings (in these cases further restrained by linguistic simplicity through child-directed speech or Simple English), much smaller architectures trained for shorter periods of time can still exhibit similar or even improved performance compared to larger models.

Apart from model architecture, also the concrete *learning* process (viz. the training goal) in current language models requires theoretical scrutiny. Whether it uses prediction in context, next word prediction or next sentence prediction, learning always involve a form of *prediction*. While prediction effects in language are well documented (for an overview, see Ryskin et al., 2020), it remains an open question whether the current flavor of prediction in language model training aligns with its cognitive counterpart. While the unidirectional prediction goal in autoregressive models (like those from the GPT-family) appears cognitively more plausible than bidirectional prediction, as employed in e.g. BERT-like models – after all, humans can only predict from what they have already processed, and not from the following (not yet perceived) contexts – other modalities of language acquisition like reading often involve explicit instruction with bi-directional prediction (e.g. fill-in-the-blank exercises).

3 Child-directed speech is tailored to children's needs

Child-directed speech differs from regular adult-adult conversation in several crucial aspects. It should be noted that its specific features[2] are not *exclusive* to child-directed speech, but rather *preferred* in this specific register. As such, child-directed speech is a gradient concept, where certain utterances stick out as more prototypical instances. We use the following four features of child-directed speech to define a prototypicality ranking that we employ in our curriculum approach.

[1] Our code can be found at `https://github.com/clause-bielefeld/gpt-wee`

[2] The following section only reiterates the features directly relevant to the current modelling task. For a more comprehensive overview across all layers of linguistic analysis, Saxton (2017) and Clark (2009, 32–41) should be consulted.

The first feature is word length. Saxton (2017) describes how the child-directed vocabulary is mostly restricted to short words grounded in the direct spatial and temporal proximity of the child. Concrete objects are favoured over abstract concepts. As Zipf (1935) already noted, word length is inversely proportional to word frequency. Furthermore, longer words have a higher informational content (Piantadosi et al., 2011) and are thus not ideal for the – still developing – linguistic and cognitive capabilities of children.

Secondly, word frequency itself, although it is a contested notion (Saxton, 2009), plays an important role in language acquisition. Apart from its role in the input, it is also reflected in children's earliest utterances, which are highly item-based and revolve around so-called pivot schemas (Braine and Bowerman, 1976; Tomasello, 2000; Diessel, 2013a), for example *more [NP]*, where *more* as the static lexical element is combined with a slot for a noun phrase. Ambridge et al. (2015) show evidence for a direct relationship between the age of acquisition of linguistic forms and their frequency in the input. Importantly, the Zipfian distribution of lexical elements in child-directed speech is stable across the development span of children as well as across typologically diverse languages (Lavi-Rotbain and Arnon, 2023). From these empirical findings, we deduce that child-directed utterances with more frequent lexical items (across the entirety of the input) can also be seen as more prototypical.

Thirdly, moving from the lexical to the syntactic level, Cameron-Faulkner et al. (2003) show that the majority of child-directed speech does not consist of canonical subject-predicate sentences, but rather of questions, imperatives and an enormous amount of fragments without a regular predicate. For different input types, these distributions vary considerably. Children's books, for example, feature a much higher amount of subject-predicate and complex sentences (with two or more lexical verbs) than ordinary speech (Cameron-Faulkner and Noble, 2013). Because the everyday child-directed input (e.g. in toyplay or meal sessions) contains more fragments compared to these specialised kinds of input, we conclude that shorter utterances are also more prototypical for child-directed speech.

Finally, Cameron-Faulkner et al. (2003) also show that the majority of child-directed utterances begin with what they call "lexical frames" – highly frequent utterance-initial, mostly two- or three-word, lexical sequences which are stable across development and different caregivers. These specific frames are thought to facilitate the acquisition of item-based constructions, which then later gradually emerge into a complete mental grammar. From this, we conclude that child-directed utterances beginning with highly frequent frames, here measured in trigrams, are also more prototypical.

As Geeraerts (1989) notes, prototype theory is prototypical in itself and not a monolithic framework. For the sake of the present analysis, we define the overall prototypicality of an utterance as the shared centrality along all axes of the mentioned prototype criteria – in concrete terms this means that we combine the utterance ranks to determine a final rank for each utterance.

4 Curriculum learning

Curriculum learning is an approach to machine learning where "the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and gradually more complex ones" (Bengio et al., 2009, 41). They propose two advantages: less training time (as the learner does not waste time on predicting noisy or hard examples too early), and an orientation into "better areas of the training space" – local minima during optimization.

This approach has been proven effective across a variety of tasks, for example in vision and language (Zhang et al., 2021) or reinforcement learning (Narvekar et al., 2020), but it remains questionable under which circumstances considerable advantages emerge. Wu et al. (2021) show that for established benchmarks, the advantages are marginal to non-existent. In contrast, the benefits are the most pronounced for problems with noisy training data. Child-directed speech, with its high amount of fragmentary utterances, can also be considered somewhat *noisy* input which, in conclusion, might benefit from a curriculum approach.

Importantly, our flavor of curriculum learning implements usage-based and cognitive principles as the source of the concrete curriculum ordering, and no engineering-based metrics, pacing functions or other kinds of transfer learning, e.g. those with teacher networks that determine the examples' difficulty (as in Hacohen and Weinshall, 2019). Due to the *a priori* nature of these aspects, we employ a vanilla approach to curriculum learning (Soviany et al., 2022), meaning that we only order the exam-

	Small	Medium	Large
Vocabulary size	4k	8k	16k
Hidden layers	2	2	4
Attention heads	2	2	4
Embedding size	64	128	256
Context size	64	128	128
Parameters	0.42M	1.55M	7.52M

Table 1: Model parameters

ples once and then provide them to the training algorithm in this static order, to maintain comparability with equivalent no-curriculum models. Interestingly, the BabyBERTa experiments implemented a somewhat comparable functionality. They showed that, in their own grammatical test suite, models benefit from this *scaffolding*, i.e. first training on child-directed speech and only later on more complex registers and non-dialogue input data.

5 Implementation

5.1 Training

As training data, we used the `babylm_10M` data set from the `strict-small` submission track for the BabyLM challenge. It consists of a mixture of child-directed and adult-directed speech, e.g. from CHILDES (MacWhinney, 2000), as well as written language, e.g. from Wikipedia. The exact composition of the corpus is described in Warstadt et al. (2023a). For evaluation during the training process, we used the `babylm_test`[3] data set.

We trained models of three different sizes, each once with and once without curriculum learning. Table 1 shows the different parameter configurations[4]. The training process was implemented in the huggingface transformers library (Wolf et al., 2020). As already mentioned, we decided on a GPT2 architecture (Radford et al., 2019) to account for the sequential nature of language. A BPE tokenizer was trained with a vocabulary size of 4k/8k/16k subword tokens. Before tokenization, all textual input was normalized in terms of capitalization and eventual diacritics. For the curriculum models, the pre-ordered examples were dynamically loaded in unshuffled batches during training time, which preserved the calculated order based on the prototypicality measures. We supplied the models with training batches of size 32. Regarding training hyperparameters, we used the cosine learning rate scheduler with a learning rate if 5e-4, weight decay of 0.1, 1k warm-up steps and 8 gradient accumulation steps. All models were trained for exactly 10 epochs in the non-curriculum setting and roughly 10 epochs in the curriculum setting, where we could not set the exact number of epochs due to the dynamic data loading. The models were evaluated after each training epoch. After those 10 epochs, the losses mostly stabilized. We did not conduct any kind of extensive hyperparameter search. Instead, we only used the default configurations for GPT-2 training, including dropout probabilities of 0.1 and layer normalization. By doing so, we tried to stay as close as possible to the vanilla configuration, which allows us to better assess the effects of smaller architectures in isolation.

The models were trained on a GPU workstation equipped with an Intel Core i7-4770 CPU (3.40GHz), 32GB of RAM and an NVIDIA GeForce GTX 1080 Ti GPU. Due to the small number of parameters, training times varied between 3–4 hours for the smallest models to 20h for the largest models.

5.2 Sentence scoring

To order the curriculum input sentences, we determined four different scores based on the aforementioned prototypicality criteria of child-directed speech. For each utterance/sentence in the training data (delimited by sentence-final punctuation or line breaks, dependening on the corpus file), we calculated the following:

- the **average word length** of a sequence, measured by the mean number of characters for all tokens in a sequence
- the **average word frequency** of a sequence, measured by taking the mean of the individual token frequencies across the whole training data
- the **utterance length**, measured as the number of lexical tokens in the sequence
- the **frame frequency**, calculated as the amount of times that the three utterance-initial tokens occur in that configuration through the training data

[3] A dev data set was also provided, but due to their equivalent size it the choice between did not affect the outcome of the training process.

[4] From this point onwards, we will denote the models by the vocabulary size of their tokenizer.

	Mean (SD)
Frame frequency	188.76 (917.04)
Utterance length	8.01 (9.21)
Mean word length	4.28 (1.37)
Mean word frequency	55153.93 (42877.18)

Table 2: Distribution of scoring variables

We operationalized the frame frequency as exactly three utterance-initial tokens because this number provides a good trade-off between the open-ended nature of sentences (and their long-tail distribution of final lexical items) and the number of fixed lexical items that certain syntactical constructions are associated with.

For each value, we calculated the respective rank of the utterance across all utterances. The final "prototypicality rank" for each utterance was calculated by taking the sum of these four ranks and then ranking by this sum.

Mean values and standard deviations for the four criteria are reported in table 2. Especially for the frame frequency and the utterance length, the distributions are heavily skewed and indicate long-tail distributions. The mean word length of approximately 4 with a standard deviation of 1.34 is to be expected, whereas the distribution of the sentences' mean word frequency also appears to be heavily skewed. As Lavi-Rotbain and Arnon (2023) show how pervasive Zipfian distributions are on a lexical level, it is not surprising that other properties of language, e.g. lexical frames, follow similar laws.

6 Results

6.1 Training

We evaluated the models after every 5k training steps during the approximately 40k training steps, returning 8 data points for training and evaluation loss. Their development is reported in appendix A (figures 1, 2 and 3). Across all models, the evaluation loss for the curriculum learning is initially much higher than the other losses, whereas the evaluation loss for the normal, randomized learning is the lowest. This is not surprising, however, as the evaluation data was not re-ordered and thus many linguistic features present in it were not yet processed by the curriculum models during earlier training steps. The regular training losses share a very similar development across all model sizes. Between the model sizes, differences are more pronounced in the later stages of training. Noticeably, the smallest model seems to converge the earliest, while the largest model might have benefited from even further training. Furthermore, the curriculum evaluation loss stays much higher for the larger model, whereas it converges in similar dimensions of the training losses for the smaller models. As such, both an effect of the curriculum learning (albeit not strictly positive) and an interaction between model size and (non-)curriculum learning can be reported.

6.2 Zero-shot evaluation with BLiMP

We tested our models with the evaluation suite supplied by the BabyLM challenge (Gao et al., 2022; Warstadt et al., 2023a), which included zero-shot evaluation tasks as well as tasks requiring additional fine-tuning. The zero-shot tasks are taken from the BLiMP evaluation suite (Warstadt et al., 2020a), which consists of minimal acceptable/unacceptable pairs of sentences across a wide variety of linguistic phenomena. To evaluate models, these sentences are scored by the models for their likelihood. A model is said to have acquired grammatical knowledge of a specific phenomenon if it consistently scores the acceptable sentences higher.

The results for the BLiMP tasks are shown in Tables 3 and 4. When comparing our own GPT-Wee models, we find that there is no straightforward effect of model size on task performance. For the majority of tasks, the performance increases with model size, whereas some tasks (e.g. hypernym, island effects) show light inverse scaling behavior. On most tasks, the effect of curriculum learning is small and rather mixed (positive or negative), when compared to the respective baseline (same model size, without curriculum). Overall, model size has a larger effect than curriculum learning. In a few task-model combinations, though, curriculum learning has a very strong positive effect (16k model/anaphor agreement, 8k model/irregular forms, 16kmodel /quantifiers) and in one case a strong negative effect(8k/NPI). Thus, if at all, it is rather the medium-sized or larger models than the small models which seem to benefit from the curriculum. For the quantifiers task, for example, the curriculum model with a 16k vocabulary outperforms all other models by approx. 18%.

Compared to the baseline results[5], we find that

[5] Taken from https://github.com/babylm/

	anaphor agreement:	argument structure:	binding:	control raising:	determiner noun agreement:	ellipsis:	filler gap:	irregular forms:	island effects:
4k	63.50	60.11	61.26	60.78	65.34	32.56	64.11	68.65	47.80
4k (cu.)	57.98	57.86	63.97	60.78	64.58	35.45	66.06	70.03	43.05
8k	71.06	64.69	65.75	62.64	78.69	44.11	62.68	82.29	42.49
8k (cu.)	64.37	63.86	65.94	62.88	75.96	44.86	65.70	90.13	37.07
16k	73.82	71.91	68.97	66.26	88.36	54.56	68.67	86.06	41.03
16k (cu.)	82.87	69.51	65.24	63.21	85.52	55.43	66.65	77.56	40.88
OPT	63.8	70.6	67.1	66.5	78.5	62	63.8	67.5	48.6
RoBERTa	81.5	67.1	67.3	67.9	90.8	76.4	63.5	87.4	39.9
T5	68.9	63.8	60.4	60.9	72.2	34.4	48.2	77.6	45.6

Table 3: Results (accuracies) of zeroshot BLiMP and BLiMP Supplement evaluation measures for our GPT-Wee models and baseline models (OPT, RoBERTa and T5)

	npi licensing:	quantifiers:	subject verb agreement:	hypernym:	qa congruence easy:	qa congruence tricky:	subject aux inversion:	turn taking:
4k	49.95	54.87	50.62	52.21	48.44	39.39	81.53	45.71
4k (cu.)	49.47	55.41	52.09	50.00	43.75	44.85	80.09	43.57
8k	52.10	60.90	56.24	49.77	51.56	32.12	82.58	50.36
8k (cu.)	37.97	60.38	57.81	49.88	50.00	40.00	85.44	46.43
16k	51.97	59.61	66.49	49.42	57.81	28.48	80.09	54.29
16k (cu.)	46.60	78.54	65.82	50.93	53.12	33.33	83.46	56.79
OPT	46.7	59.6	56.9	50.0	54.7	31.5	80.3	57.1
RoBERTa	55.9	70.5	65.4	49.4	31.3	32.1	71.7	53.2
T5	47.8	61.2	65.0	48.0	40.6	21.2	64.9	45.0

Table 4: Results (accuracies) of zeroshot BLiMP and BLiMP Supplement evaluation measures for our GPT-Wee models and baseline models (OPT, RoBERTa and T5), contd.

our smaller models do not perform considerably worse on average, and outperform the baseline models for selected tasks. For example, a few of our small models are surprisingly good at island effects, hypernyms, qa congruence, or subject-auxiliary inversion. As the baseline results are derived from BERT/OPT/T5 models with much larger architectures and higher parameter numbers (e.g. 125M parameters for the OPT model, with 12 hidden layers, 12 attention heads, a 50k token vocabulary and intermediate embeddings of size 768), we are pleasantly surprised by the comparatively good results which our models achieve.

6.3 (Super)GLUE and MSGS evaluation

For the evaluation tasks requiring additional fine-tuning, we only collected results for our submitted, medium-sized curriculum model due to constraints in computing power and time.

The GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks involve fine-tuning on a variety of tasks, e.g. question answering, correct identification of entailment or the extraction of correct co-references. As such, this benchmark is more focused on semantic and pragmatic aspects.

Regarding the (Super)GLUE scores (table 5), a similar picture to the BLiMP scores emerges. Across many of the tasks, our model performs in similar ranges as the baselines, often better than the T5 baseline and more similar to the OPT baseline. Although our models are considerably smaller, they seem to provide similar starting points for fine-tuning on additional data.

Finally, the Mixed Signals Generalization Set (MSGS) introduced by Warstadt et al. (2020b) also contains different ambiguous binary classification tasks. The test sentences are ambiguous in the sense of allowing both surface generalizations and generalizations that require deeper linguistic understanding of structure. Additionally, control experiments are included that test whether a feature is actually encoded. The scores reported in table 6 are correlations, where a value greater than zero denotes a preference for linguistics generalizations, and a value below zero shows a preference for surface generalizations. The performance of our model is (once more) very similar to the baselines. The control tasks show that our model does encode the tested features, but the test tasks show a system-

evaluation-pipeline

	CoLA (MCC)	SST-2	MRPC (F1)	QQP (F1)	MNLI	MNLI-mm	QNLI	RTE	BoolQ	MultiRC	WSC
8k (cu.)	4	80	82	66	60	61	61	60	61	55	60
Maj.	0.0	50.2	82.0	53.1	35.7	35.7	35.4	53.1	50.5	59.9	53.2
OPT	15.2	81.9	72.5	60.4	57.6	60.0	61.5	60.0	63.3	55.2	60.2
RoB.	25.8	87.0	79.2	73.7	73.2	74.0	77.0	61.6	66.3	61.4	61.4
T5	11.3	78.1	80.5	66.2	48.0	50.3	62.0	49.4	66.0	47.1	61.4

Table 5: (Super)GLUE scores (accuracies unless otherwise stated as MCC or F1) for our 8k curriculum GPT-Wee model, the majority baseline and the three provided model baselines

	CR (Control)	LC (Control)	MV (Control)	RP (Control)	SC (Control)	CR_LC	CR_RTP	MV_LC	MV_RTP	SC_LC	SC_RP
8k (cu.)	43	93	37	100	76	0	-74	-99	-99	-57	-73
OPT	50.8	53.6	99.5	99.9	77.2	0.4	-70.3	-72.1	-77.6	13.8	-68.9
RoB.	43.1	100.0	97.7	76.7	86.2	-28.3	-77.7	-99.3	-79.4	16.3	-45.0
T5	21.1	100.0	33.4	82.5	77.6	-78.3	-62.0	-100.0	-79.7	-25.3	-39.4

Table 6: MSGS scores (MCC) for our 8k curriculum GPT-Wee model and the three provided model baselines

atic bias for surface generalizations. However, this behavior is also (with minor deviations) observable in the baseline models. All naïve models fail to generalize based on the linguistic features.

6.4 Age-of-acquisition evaluation

Additionally, the BabyLM evaluation suite provided an age-of-acquisition evaluation (Portelance et al., 2023). The calculated scores (table 7) are Mean Absolute Deviation (MAD) values, measured in months, representing the difference between the actual average age-of-acquisition (AoA) of the tested word among American English-speaking children and the predicted AoA based on our models' average surprisal scores. Lower MAD scores indicate better performance. We calculate these scores for all of our models and find that the individual differences between the models are tiny or nonexistent, and roughly the same as the baseline results provided by the challenge. As such, also here the effect of the choice of a specific language model architecture does not seem to have much of an influence on the evaluation metric.

7 Discussion

The present analysis set out to investigate the influence of a usage-based factors, input ordering, and an architectural factor, model size, on the learning processes (and successes) of language models. We found that both factors have a certain influence on the training process and the model performance. While model size affects the performance in linguistic evaluation, the effect is not linear across tasks. For zero-shot tasks, the majority show improved scores, although a few scores decrease with increasing model size. Compared to much larger baseline models, our models' performance is not considerably worse. Especially the non-linear effects of model size warrant further inspection: it remains unclear which internal factors (context length, vocabulary size, model parameters, number of training epochs, etc.) contribute to which developments, and how these factors interact with each other. For the tasks requiring additional fine-tuning, our 8k curriculum model also performed similarly to the baselines. Especially for the (SUPER)Glue benchmark, a more semantics- and pragmatics-oriented benchmark, the performance was quite in line with the baseline models, hinting at the acquisition of a fair amount of the needed information. The MSGS benchmark, however, showed that our model systematically picks up surface generalizations. Yet, this also applies to the much larger baselines.

The usage-inspired naïve ordering approach to curriculum learning also has no straightforward effects on model performance. Especially during the training process, differences to traditional, randomized learning are observable. Although it appears to be somewhat detrimental to overall performance, certain specific evaluation tasks are positively influenced. The results thus remain inconclusive. From a usage-based viewpoint, Diessel (2013b) stresses the importance of deictic pointing and joint attention as (extralinguistic) language acquisition fac-

	Overall	Nouns	Predicates	Function words
4k	2.07	2.00	1.84	2.65
4k (cu.)	2.06	1.99	1.84	2.64
8k	2.07	2.00	1.82	2.65
8k (cu.)	2.06	2.00	1.82	2.64
16k	2.06	2.00	1.83	2.65
16k (cu.)	2.06	2.00	1.83	2.58
OPT	2.03	1.98	1.81	2.57
RoBERTa	2.06	1.99	1.85	2.65
T5	2.04	1.97	1.82	2.64

Table 7: MAD scores between actual AoA and the predicted AoA, for our GPT-Wee models and the three baselines

tors. Besides, also intention reading, role reversal and imitation (Tomasello, 2003, 21–28) are important acquisition factors that LLMs cannot mirror – they are strictly confined to statistical/frequency-driven aspects of usage-based theory (which are nevertheless very important, as noted by Ambridge et al., 2015). Still, we only have child-directed *speech* for training, and no real child-directed *communication*, which connects speech with such extralinguistic factors and influences utterance prototypicality beyond the modalities that we were able to include in the present experiment.

The non-improvements added by the curriculum approach also further add to the debate on what language models mean for linguistic theory. For example, Pannitto and Herbelot (2022) and Piantadosi (2023) have stressed the anti-Chomskyan evidence provided by the successes of language models. Curriculum learning looks like an obvious choice when trying to implement usage-based findings in the training process for (smaller) language models. However, this does not seem to work with the simple form of curriculum learning based on prototypicality measures that we used in this paper. For that, several explanations are possible: 1) more advanced curriculum approaches are needed, with different and more directed ways of ordering and optimizing the curriculum, 2) curriculum learning may not be the right choice for small models (it seems that, if at all, it was rather the larger models which showed tendencies of improvement. Also, other options for implementing usage-based accounts might just work better (e.g. models with dynamic structures and growing numbers of nodes). After all, real human neural networks grow and mature while they are constantly shaped and re-shaped by linguistic input and processing. As such, it also remains hard to interpret language models, their parts and their performance on various evaluation suites in a coherent way. The integration of more linguistic factors into the training process needs to be tested in this regard. For example, Yehezkel and Pinter (2023) propose a subword tokenization algorithm that incorporates contextual information and creates vocabularies that seem to align more with classical ideas of morphology. It remains an open question whether such alterations and other linguistic experiments in the training process would also improve the linguistic quality of the generated output.

8 Conclusion

The BabyLM challenge set out to test different approaches to language modelling with small data. When looking at the leaderboard[6], we find that our model is located in the lower section of the rankings. However, the best-performing models implement much more complex learning strategies and larger architectures. We, on the other hand, decided on very small architectures. As such, our results can be seen as a success: benchmark performance seems to be much more strongly constrained by the concrete linguistic make-up of the training data and not so much by model size alone, as our downsizing apprach shows. This also confirms earlier findings from BabyBERTa (Huebner et al., 2021) and TinyStories (Eldan and Li, 2023). Our key takeaway is that a *one size fits all* approach to language model architectures should not be adopted without further thought, and that training data quality and make-up should be valued more. Besides, we also tested a usage-based approach to curriculum learning. Although our curriculum models are generally not superior to the regular, randomized

[6] At https://dynabench.org/babylm

models, some zero-shot evaluation tasks did benefit from it. Additionally, small model size and the curriculum training did not have a detrimental effect on pre-training for the tasks that require fine-tuning. Still, our results show that a much more fine-grained approach to the evaluation of such factors is needed. As language model engineers, we can choose between a large variety of evaluation suites that test along all levels of linguistic analysis and across many different task set-ups. However, we do not know how the changes in individual, low-level variables (e.g. number of hidden layers, context size) impact specific factors of linguistic performance (e.g. the ability to judge acceptability for island effects, or the ability to correctly predict entailment). To correctly interpret such choices, further systematic analyses are clearly needed.

Acknowledgements

We acknowledge financial support by the project "SAIL: SustAInable Life-cycle of Intelligent Socio-Technical Systems" (Grant ID NW21-059A), which is funded by the program "Netzwerke 2021" of the Ministry of Culture and Science of the State of Northrhine Westphalia, Germany.

References

Ben Ambridge, Evan Kidd, Caroline F. Rowland, and Anna L. Theakston. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, 42(2):239–273.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, Virtual Event Canada. ACM.

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, Montreal Quebec Canada. ACM.

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

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

Thea Cameron-Faulkner and Claire Noble. 2013. A comparison of book text and Child Directed Speech. *First Language*, 33(3):268–279.

Eve V. Clark. 2009. *First Language Acquisition*, 2nd ed edition. Cambridge University Press, Cambridge ; New York.

Holger Diessel. 2013a. 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. 2013b. Where does language come from? Some reflections on the role of deictic gesture and demonstratives in the evolution of language. *Language and Cognition*, 5(2-3):239–249.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English?

Jeffrey L. Elman. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99.

Jeffrey L. Elman, Elizabeth L. Bates, Mark H. Johnson, Annette Karmiloff-Smith, Domenico Parisi, and Kim Plunkett. 1996. *Rethinking Innateness: A Connectionist Perspective on Development*. Neural Network Modeling and Connectionism. MIT Press, Cambridge, MA.

Leo Gao, Jonathan Tow, Stella Biderman, Charles Lovering, Jason Phang, Anish Thite, Fazz, Niklas Muennighoff, Thomas Wang, Sdtblck, Tttyuntian, Researcher2, Zdeněk Kasner, Khalid Almubarak, Jeffrey Hsu, Pawan Sasanka Ammanamanchi, Dirk Groeneveld, Eric Tang, Charles Foster, Kkawamu1, Xagi-Dev, Uyhcire, Andy Zou, Ben Wang, Jordan Clive, Igor0, Kevin Wang, Nicholas Kross, Fabrizio Milo, and Silentv0x. 2022. EleutherAI/lm-evaluation-harness: V0.3.0. Zenodo.

Dirk Geeraerts. 1989. Introduction: Prospects and problems of prototype theory. *Linguistics*, 27(4):587–612.

Guy Hacohen and Daphna Weinshall. 2019. On the power of curriculum learning in training deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544. PMLR.

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.

Ori Lavi-Rotbain and Inbal Arnon. 2023. Zipfian Distributions in Child-Directed Speech. *Open Mind*, 7:1–30.

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

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E. Taylor, and Peter Stone. 2020. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(1).

Ludovica Pannitto and Aurelie Herbelot. 2022. Can Recurrent Neural Networks Validate Usage-Based Theories of Grammar Acquisition? *Frontiers in Psychology*, 13:741321.

Steven T. Piantadosi. 2023. Modern language models refute Chomsky's approach to language.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal.

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.

Rachel Ryskin, Roger P. Levy, and Evelina Fedorenko. 2020. Do domain-general executive resources play a role in linguistic prediction? Re-evaluation of the evidence and a path forward. *Neuropsychologia*, 136:107258.

Matthew Saxton. 2009. The Inevitability of Child Directed Speech. In Susan Foster-Cohen, editor, *Language Acquisition*, pages 62–86. Palgrave Macmillan UK, London.

Matthew Saxton. 2017. *Child Language: Acquisition and Development*, 2nd edition edition. SAGE, Los Angeles.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum Learning: A Survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Kenneth O. Stanley and Risto Miikkulainen. 2002. Evolving Neural Networks through Augmenting Topologies. *Evolutionary Computation*, 10(2):99–127.

Michael Tomasello. 2000. 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.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for Papers – The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023b. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

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

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. 2021. When do curricula work? In *International Conference on Learning Representations*.

Shaked Yehezkel and Yuval Pinter. 2023. Incorporating context into subword vocabularies. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 623–635, Dubrovnik, Croatia. Association for Computational Linguistics.

Jiwen Zhang, zhongyu wei, Jianqing Fan, and Jiajie Peng. 2021. Curriculum learning for vision-and-language navigation. In *Advances in Neural Information Processing Systems*.

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

A Training losses

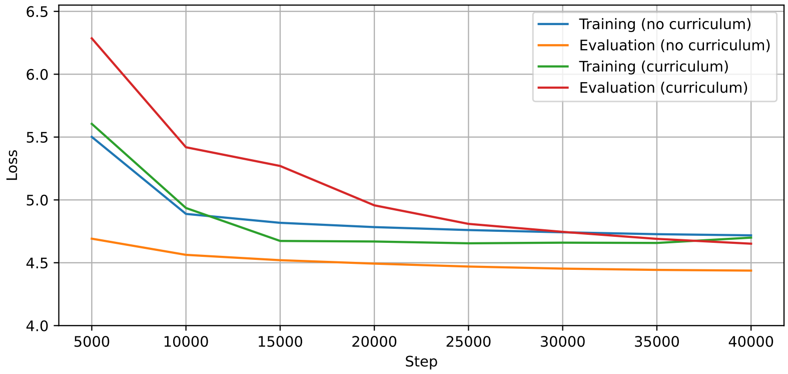

Figure 1: Training and evaluation losses for the 4k vocabulary models, calculated every 5k steps

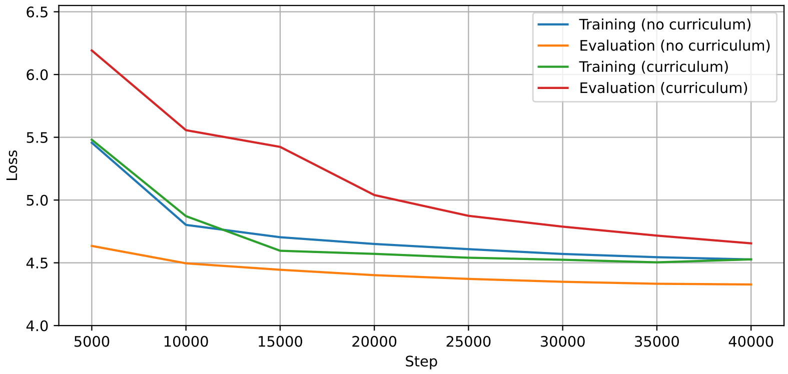

Figure 2: Training and evaluation losses for the 8k vocabulary models, calculated every 5k steps

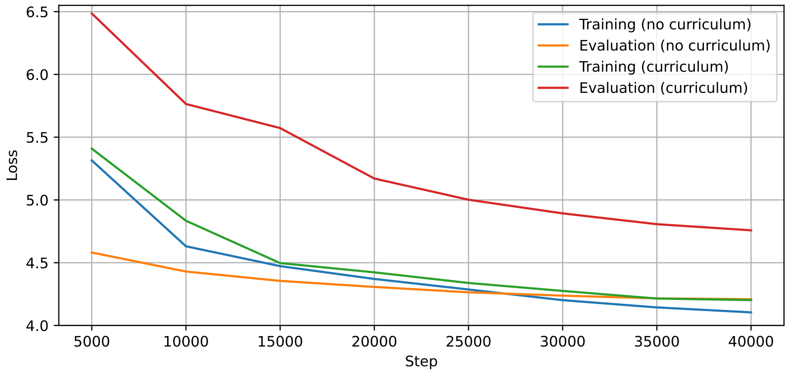

Figure 3: Training and evaluation losses for the 16k vocabulary models, calculated every 5k steps