Graphemes vs. phonemes: battling it out in character-based language models

Bastian Bunzeck, Daniel Duran, Leonie Schade and **Sina Zarrieß**
CRC 1646 – Linguistic Creativity in Communication
Department of Linguistics
Bielefeld University, Germany
{firstname.lastname}@uni-bielefeld.de

Abstract

We present grapheme-llama and phoneme-llama, character-based language models trained for the 2024 BabyLM challenge. Through these models, we explore an under-researched approach to downsizing: replacing subword-based tokenization with character-level tokenization, drastically reducing the vocabulary size. The grapheme model is trained on a standard BabyLM dataset, while the phoneme model uses a phoneme-converted version of this dataset. Results show that grapheme-based models perform better overall, achieving scores comparable to subword-based models on grammatical benchmarks. Despite lower performance, phoneme models also demonstrate promising grammatical learning. We argue that our results challenge conventional wisdom on language modeling techniques and open up novel research questions with character- and phoneme-based models as objects of inquiry.

1 Introduction

While *large* language models continue to beat benchmarks, their parameter numbers, amounts of training corpora and training FLOPs are ever increasing. More recently, however, a new research focus on ecologically friendly, data-efficient and possibly cognitively plausible language models – so called BabyLMs – has emerged. But what makes a language model a *BabyLM*? For the BabyLM challenges (Warstadt et al., 2023; Choshen et al., 2024), BabyLMs are defined by extremely constrained data settings. In this constrained data setting, the best scoring models in the 2023 challenge employed highly sophisticated and large-ish architectures: ELC-BERT (Charpentier and Samuel, 2023) used numerous architectural improvements over standard encoders, while BabyLlama (Timiryasov and Tastet, 2023) was distilled from various larger teacher models. Models with architectures downsized similarly to their training data (e.g. by Veysel Çağatan, 2023, Bunzeck and Zarrieß, 2023 or Fields et al., 2023) did not fare as well on standard benchmarks.

As our submission to the 2024 BabyLM challenge (Choshen et al., 2024), we present grapheme-llama[1] and phoneme-llama[2]. We replace the standard subword-based tokenization algorithms with naive character-based tokenization, leading to a drastic decrease in vocabulary size. We show that when such simplifications are combined with state-of-the-art architectures like Llama (Touvron et al., 2023b), the resulting models still achieve considerable grammatical proficiency and provide useful inductive biases for further fine-tuning. While the grapheme model is trained on the standard 100M BabyLM data, our phoneme model is trained on a version of this data set converted to phonemes[3]. Although it performs generally worse than its grapheme counterpart, the phoneme model still manages to learn the grammatical phenomena in a matched BLiMP data set quite well. In the light of these results, we offer some discussion points for phoneme-based language modeling, the pitfalls it is currently facing and its general potential. In sum, we argue that these results open fruitful avenues for further research on small language models and question "common wisdom" in current language modeling practices.

2 Related work

Small LMs/downsizing: Recently, there has been a surge in interest in small-ish language models. The arguably first BabyLM, BabyBERTa

[1] https://huggingface.co/bbunzeck/grapheme-llama

[2] https://huggingface.co/bbunzeck/phoneme-llama

[3] In line with the G2P literature (cf. Moore and Skidmore, 2019; Ashby et al., 2021), we use (i) the term "phoneme" loosely to refer to (symbols for) types of speech sounds and (ii) the term "grapheme" loosely to refer to the letters of orthographic alphabets.

(Huebner et al., 2021), followed a combined (i.e. data *and* architecture) downsizing approach and showed that dramatically less training data can result in remarkable linguistic proficiency with a small model architecture. On the other hand, current "small" models often employ more complex strategies to achieve compactness, e.g. distillation with teacher and student models (Timiryasov and Tastet, 2023), or reduction of number precision (Wang et al., 2023). These models' "smallness" is only achieved after complex training procedures. In contrast to these developments, the BabyLM 2023 submissions by Veysel Çağatan (2023), Bunzeck and Zarrieß (2023) and Fields et al. (2023) used *a priori* small models (in terms of parameter size) to show the lower bounds of knowledge learnability from small data. They all showed that very small models (even models with a parameter size below 1M) can achieve scores equal to much larger baselines on standard evaluation tasks like BLiMP or GLUE. As such, these successful experiments give impetus for our current models: against common wisdom, the reduction of certain models hyperparameters does not have to have a detrimental effect on performance (a fact also corroborated by Muckatira et al., 2024). Comparable studies have neither focused on character-level tokenization nor on phoneme-based representations (see paragraphs below for the most comparable studies available), so we pioneer into this uncharted territory with our models.

Character-level LMs: While research on LMs with character-level tokenization is not exactly scarce, they have yet to gain widespread adoption. Character-based models have been inplemented for different architectures: the Canine (Clark et al., 2022) architecture is a character-level encoder, the ByT5 (Xue et al., 2022) models employ a T5 encoder-decoder architecture with a Byte-level tokenizer and the Charformer models (Tay et al., 2022) use a tokenization module (GBST) that learns latent subword representations from characters. For all three models it has been shown that their specific pre-training regimens do provide useful inductive biases for further fine-tuning and that such are more robust to character-level noise than regular subword-tokenization models. Moreover, phonological categories like consonants and vowels are retrievable from Canine (see Agirrezabal et al., 2023) – properties of language that are by design not captured by coarse-grained subword representations. From a more application-driven standpoint, El Boukkouri et al. (2020) have shown that character-level modeling can improve performance in the medical domain. Finally, Edman and Bylinina (2023) showed in the context of last year's BabyLM challenge that first training on a character-level vocabulary and then expanding it to the subword-level provides mixed effects on model performance, depending on the context size. It should also be noted that there are further approaches to language modeling without complex tokenization algorithms: Rust et al. (2023) show that LMs trained on pixel-based representations can help LMs excel at various syntactic and semantic tasks in typologically diverse languages, including non-Latin scripts.

Phoneme LMs: So far, phoneme-based LMs have mostly been trained as encoders to provide inductive biases for further fine-tuning on downstream tasks. PhonemeBERT (Sundararaman et al., 2021), Mixed-Phoneme BERT (Zhang et al., 2022) and XPhoneBERT (Nguyen et al., 2023) are examples for such models, which have been reported to improve downstream performance on various tasks, e.g. on text-to-speech. In contrast, the CharsiuG2P model (Zhu et al., 2022) is an encoder-decoder architecture explicitly pre-trained for grapheme-to-phoneme conversion (G2P). Purely autoregressive phoneme models have not received scientific attention, yet.

3 Methodology

3.1 Data

We train our models on the 100M BabyLM 2024 data set. This data set contains both (transcribed) spoken and written language. It includes spoken language from CHILDES (MacWhinney, 2000), the BNC (Burnard, 2007), Switchboard (Stolcke et al., 2000) and OpenSubtitles (Lison and Tiedemann, 2016), and written language from children's books in Project Gutenberg (Gerlach and Font-Clos, 2020) as well as a portion of the Simple English Wikipedia. Because the raw data contains extensive metadata and markup, we used an expanded version of the cleaning script from Timiryasov and Tastet (2023) to clean the data.

For our phoneme-based models, we then convert the cleaned data from graphemes to phonemes – a mapping from orthographic letters to sound-symbols to represent the pronunciation of the text. To convert our text to IPA (International Phonetic Association, 1999) symbols, we use the rule-based G_i2P_i system for G2P-conversion (Pine et al., 2022)[4], expanded by a manual replacement list that we compiled for contractions that this tool does not handle well. As the authors report no G2P accuracy for English, we conduct a manual evaluation on three short texts. We find a word-error-rate of 5.8% (tokens=363, errors=21), which we deem as sufficient for the sake of the current paper. For evaluation purposes, we also perform the same G2P conversion on the BLiMP data. We make this data set[5] and our converted training data[6] available on the Hugging Face hub.

3.2 Training

We use the `transformers` library (Wolf et al., 2020) to train four small, character-level llama models (Touvron et al., 2023b). All our models share equivalent model internals and training hyperparameters:

- Training tokens: 100M
- Hidden layers: 8
- Attention heads: 8
- Embedding size: 512
- Context size: 64
- Number of parameters: 15M/14.9M (grapheme-based/phoneme-based models)

We train two models on the original grapheme-based BabyLM data and two models on our converted phoneme-based data: for each data regimen, one model with whitespaces separating lexical tokens and one without these whitespaces. As we experiment with removing information about words by not using sub-word tokenization, the models without whitespaces can be seen as more extreme variants of the same training setting – they have (apart from beginning and end of sequences) no access to word segmentation information at all. To force the models to use more local information, we restrict the context size to 64 tokens (although we acknowledge that this might lead to detrimental performance on tasks that require longer contexts, especially EWoK and GLUE).

To implement character-level language modeling, we modify the tokenizers used for our models. Instead of the standard BPE tokenization algorithm, we simply fill our tokenizers' vocabularies with all unique characters in the respective pre-training corpora. For the grapheme-based models, this adds up to a vocabulary size of approx. 360. For the phoneme models, the vocabulary size is approx. 260. Next to the standard ASCII and IPA characters, these vocabularies are still so "large" due to a number of emojis and other non-linguistic Unicode characters included. Because some IPA symbols are also ordinary letters of Latin alphabets, and also due to the aforementioned non-alphabetic symbols, the vocabularies of the models share 118 tokens.

As training hyperparameters, we chose a batch size of 16, 200 warmup steps, and a learning rate set to 3e-4 in accordance with Touvron et al. (2023a). We train our models for five epochs, equaling roughly 25–28 hours of per-model training time on a single NVIDIA RTX A4000 GPU.

3.3 Model evaluation

In line with the BabyLM challenge, we evaluate our models through the BabyLM evaluation pipeline (Choshen et al., 2024; Gao et al., 2023). It includes three tasks – BLiMP (Warstadt et al., 2020), EWoK (Ivanova et al., 2024) and (Super)GLUE (Wang et al., 2018, 2019).

BLiMP is a collection of minimal pairs (ungrammatical vs. grammatical sentences) for English, including mostly (morpho)syntactic phenomena, but also semantic and (in the supplementary data) discourse-pragmatic minimal pairs. Although it suffers from a few shortcomings (partially nonsensical sentences, cf. Vazquez Martinez et al., 2023; a too restrictive binary notion of grammaticality that does not allow creative language use, etc.), it is a valuable resource and basically *the* linguistic benchmark for the evaluation of language models. If a model consistently manages to score the grammatical sentence as more plausible (i.e. through lower perplexity) it is said to have mastered the corresponding phenomenon. We evaluate all of our models on the regular BLiMP, and additionally on a matched BLiMP that contains the BLiMP data converted to match the data set the respective model was trained on (grapheme/phoneme, whitespace/no whitespace).

EWoK (Ivanova et al., 2024) is a benchmark that is supposed to measure world knowledge by testing models on their ability to match target texts with plausible/implausible contexts. It covers domains such as material properties, physical dynamics or

[4] We also tried a neural system (Zhu et al., 2022), but found it to be much less performant and of slightly worse transcription quality.

[5] https://huggingface.co/datasets/bbunzeck/phoneme-blimp

[6] https://huggingface.co/datasets/bbunzeck/phoneme-babylm-100M

social interactions. The sentence pairs function as minimal pairs (of pairs) and can therefore be evaluated in the same way as BLiMP examples. As both our grapheme models and the BabyLM baselines do not perform above chance on this benchmark, we decided not to create a phoneme version.

The (Super)GLUE tasks (Wang et al., 2018, 2019) are focused on more fine-grained language understanding and involve additional fine-tuning on task examples. As such, they measure how well our pre-training procedure supplies our models with useful inductive biases for the acquisition of these reasoning tasks, e.g. textual entailment or sentiment prediction. For reasons of time and resources, we opted to do parameter-efficient fine-tuning with LoRA (Hu et al., 2022) instead of full fine-tuning runs. In contrast to the provided fine-tuning script, we opted for only 16 epochs and a larger learning rate of 5e-4, in hopes to help our models converge faster. Due to a technical problem (and lack of time), we could only run one fine-tuning epoch for the MNLI sub-task. We also opted to not create a phonemized (Super)GLUE data set, for the same reasons as for EWoK.

4 Results

4.1 Zero-shot

The BLiMP results are collected in Table 1. With regard to the standard grapheme and whitespace BLiMP, the corresponding grapheme model also performs best. With a score of almost 72%, our character-based grapheme model is close to the subtoken-based autoregressive baseline (BabyLlama, 73.1%), and beats the masked LM baseline (LTG-BERT, 69.2%; not listed in Table 1). While the model trained without whitespace performs worse, the score of 59.88% is still far above chance. The phoneme models, on the other hand, only achieve scores that oscillate somewhat around the chance baseline. This is not surprising, as the overlap in vocabulary between the grapheme and phoneme models is small – the phoneme models can hardly retrieve any useful information from grapheme input. On the BLiMP supplement, none of our models achieve a score significantly higher than the chance baseline.

When considering the matched BLiMP evaluations, where we preprocess the BLiMP data in the same way as the pre-training corpus data, we can report much higher BLiMP scores. All four models perform way above chance, although both the G2P conversion and the deletion of all whitespace have a detrimental effect on the scores. Interestingly, the grapheme model without whitespaces achieves the best score on the BLiMP supplement (56.28%), although we can only speculate as to why (see Discussion for an attempt at explanation).

This picture gets even more complicated when we consider the individual BLiMP paradigms. The full BLiMP scores for the matched evaluation can be found in Appendix A. While the grapheme whitespace model generally performs best across the most paradigms, each model still features some high scores. For certain, highly-specific phenomena (e.g. `sentential_negation_npi_scope_filtered`), the non-whitespace phoneme model – our overall weakest model – outperforms all other models. It remains open to further inquiry whether these scores are only training noise or caused by specific linguistic factors only instantiated by this specific combination of data preprocessing steps.

The evaluation results for EWoK (Table 2) display a very uniform picture. No model achieves any considerable score above the chance baseline for any phenomenon. This is also in line with the results of the baseline models, which seemingly do not learn any "world knowledge", as measured by EWoK.

4.2 Fine-tuning

The (Super)GLUE scores can be found in Table 3. They follow no clear pattern. While the average scores for the models are rather similar (and all fairly low in comparison to the baselines, like 63.3% for BabyLlama), the scores for the individual tasks are highly varied. While the standard grapheme model achieves the highest scores on six out of eleven included tasks, all other models also get at least one highest score. Averaged across all tasks, the grapheme model without whitespace is even better than its normal counterpart. The differences between models are immense and no structured conclusions about presumed effects of any variable (grapheme/phoneme, whitespace/no whitespace) can be drawn. It is especially surprising that the phoneme models, which do not contain the full grapheme-model vocabulary and therefore sometimes lead to somewhat corrupted/distorted tokenized versions of the data (e.g. through missing tokens), still seem to impart quite useful inductive biases for many of the included sub-tasks in (Super)GLUE: Only for CoLA, MNLI and MNLI-

BLiMP version	Grapheme model	Grapheme model, no whitesp.	Phoneme model	Phoneme model, no whitesp.	BabyLlama
BLiMP	**71.69%**	59.88%	44.05%	54.02%	73.1%
BLiMP supplement	52.30%	50.12%	**55.04%**	44.47%	60.6%
Matched BLiMP	**71.69%**	68.88%	66.90%	64.88%	73.1%
Matched BLiMP supplement	52.30%	**56.28%**	55.42%	54.13%	60.6%

Table 1: BLiMP accuracies for our four models and BabyLlama baseline (random baseline = 50%)

EWoK subtask	Grapheme model	Grapheme model, no whitesp.	Phoneme model	Phoneme model, no whitesp.	BabyLlama
agent-properties	49.46%	49.68%	**50.23%**	50.05%	-
material-dynamics	49.22%	49.61%	**49.87%**	48.87%	-
material-properties	48.24%	50.00%	50.00%	**50.59%**	-
physical-dynamics	48.33%	**51.67%**	50.00%	50.00%	-
physical-interactions	47.84%	50.18%	50.18%	**51.44%**	-
physical-relations	50.73%	49.14%	49.63%	**51.22%**	-
quantitative-properties	50.96%	**52.55%**	49.36%	49.04%	-
social-interactions	49.66%	50.34%	**51.02%**	**51.02%**	-
social-properties	**51.52%**	48.78%	50.30%	48.17%	-
social-relations	49.68%	49.29%	**50.00%**	**50.00%**	-
spatial-relations	46.73%	46.33%	**51.43%**	50.20%	-
Average	49.30%	49.80%	**50.20%**	50.10%	52.1%

Table 2: EWoK accuracies for our four models and BabyLlama baseline (random baseline = 50%)

mm, the scores achieved by the (in theory unfitting) phoneme models are close or equal to the random chance baseline. For the other tasks, especially SST2 and MRPC, scores are well above chance. Here, it remains questionable whether the inductive biases of our phoneme models actually affect the performance on (Super)GLUE, or if the whole fine-tuning process equals the adoption of some heuristic shortcuts to solve the problems tested by (Super)GLUE (see Gururangan et al., 2018; Belinkov et al., 2019 for discussions of artifacts in NLI data), to which only CoLA, MNLI and MNLI-mm are robust enough to resist.

5 Discussion

General remarks: There are two commonly presented arguments against character-level tokenization (e.g. presented in Clark et al., 2022): (i) such models achieve subpar results on evaluations; and (ii) as the computational complexity of a transformer grows quadratically with the input size, the token increase yields inefficient models. To (i) we can only reply that our results speak for themselves. The strong performance of such a small Llama model on BLiMP shows that character-based models are able to learn the structure of a language as well as its subword-based sister models. The comparatively lower performance on fine-tuning tasks is likely caused by the small architecture, and could be improved with more parameters. Also, the small context size of our models might be a limiting factor for the fine-tuning tasks (and also the zero-shot EWoK evaluation, as it contains fairly long sentences). To (ii) we can reply that this is not such a big concern, as we use small models and small-ish context sizes, anyway. While this approach might not be sufficient for models with billions of parameters, it surely is for BabyLMs.

Graphemes vs. phonemes: The comparison between our grapheme and phoneme models undoubtedly concludes with a win for the grapheme models. Across all benchmarks, they outperform the phoneme models on average. No clear tendencies spring to mind when analyzing the detailed results – however, all four models achieve best scores on some sub-tasks in benchmarks. Separating noise from signal in these results remains an open task for future studies. As of now, we can only speculate why the phoneme models perform *this* worse. An easy explanation could be the absence of punctuation in phoneme models. As dots, commas and other punctuation marks perform important semantic functions in texts (see Crystal, 2015), their absence quite possibly has a negative effect on the acquired grammatical system of a language model.

Another problem could lie in the quality of our G2P system. Alphabetic writing systems generally associate letters to sounds, and vice versa. However, especially for English, the correspondences between *graphemes* and *phonemes* are not trivial and (can seem) arbitrary (Pulgram, 1951; Venezky, 1967; Emerson, 1997; Roca, 2016). Graphemes are arranged according to orthographic conventions which usually do not directly reflect a language's underlying phonological system. Grapheme-to-

GLUE subtask	Grapheme model	Grapheme model, no whitesp.	Phoneme model	Phoneme model, no whitesp.	BabyLlama
CoLA (MCC)	**0.098**	0.0668	0.0325	0	-
SST-2	**74.31%**	74.08%	69.27%	72.94%	-
MRPC (F1)	79.75%	80.62%	81.05%	**81.29%**	-
QQP (F1)	66.54%	**71.04%**	62.40%	59.57%	-
MNLI	**52.59%**	50.15%	46.92%	45.60%	-
MNLI-mm	**51.32%**	50.24%	47.40%	46.30%	-
QNLI	59.26%	**63.84%**	55.01%	52.82%	-
RTE	44.60%	43.17%	51.08%	**58.27%**	-
BoolQ	64.46%	64.65%	**64.89%**	63.85%	-
MultiRC	**57.63%**	56.23%	57.26%	57.59%	-
WSC	61.54%	61.54%	59.62%	**62.46%**	-
Average	56.50%	**56.60%**	54.40%	54.70%	69.0%

Table 3: (Super)GLUE results for our models and BabyLlama baseline

phoneme conversion, as the computational attempt to solve this problem, cannot be considered as solved. Relatively high error rates of G2P tools are still an issue in speech and language processing. For example, the SIGMORPHON shared tasks on "multilingual grapheme-to-phoneme conversion" (Gorman et al., 2020; Ashby et al., 2021; McCarthy et al., 2023) use the metrics *word error rate* (WER) and *phone error rate* (PER) for evaluation. Word error rates of the best submissions in 2020 range from 24.89 (for Georgian) to 0.89 (for Vietnamese) (Gorman et al., 2020). As such, it might be more sensible to train on manually transcribed speech. Unfortunately, such corpora are small and rare, although it might be interesting to see whether some variation in phoneme data can influence performance on standard benchmarks.

Additionally, it remains questionable how phoneme data should be represented for language modeling. Splitting a transcription into a sequence of characters for character-level tokenization introduces some issues: Unicode defines IPA base symbols as individual characters. Some diacritics (which add information on fine phonetic detail to base symbols) are defined as "Spacing Modifier Letters", others as "Combining Diacritical Marks". Thus an aspirated alveolar plosive [tʰ] or a long vowel [aː] are treated as two characters, while, depending on the treatment of composed Unicode characters, a de-voiced alveolar fricative [z̥] or a raised vowel [a̝] may be treated as one. Affricates (combined sounds), for example, may be represented as a sequence of two characters joined by a double diacritic [d͜ʒ], or as a single ligature [ʤ].

Whitespace: Finally, the detrimental effect of whitespace removal also deserves explanation and discussion. Whitespace encodes important linguistic information about word boundaries (or approximations thereof) -- information which is not available in spoken language (there, pauses between stretches of connected speech serve different purposes). Instead, prosody (e.g. word stress or intonation), provides cues to segmentation at different levels of linguistic abstraction (like words and phrases). This is, apart from whitespace, not reflected in orthographic texts and also often missing from phonetic transcriptions[7]. As such, data without whitespaces is a developmentally/cognitively/linguistically more plausible form of input. As this added plausibility comes with the loss of information, it is not surprising that scores for non-whitespace models are generally lower. A notable exception is the high score of the non-whitespace grapheme model for the matched BLiMP supplement. This might be a side effect of our very small context size. The BLiMP supplement contains inter alia dialogue phenomena with long dependencies. The models without whitespace can take in more (non-whitespace) characters, and in the light of our rather small context size, it might be the case that the whitespace models cannot process enough information to actually grasp these phenomena.

6 Conclusion

This paper has shown two things: (i) character-based tokenization is a viable alternative for small language models and (ii) phoneme-based LMs can also perform reasonably well on common benchmarks, although grapheme models are superior. With the drawbacks (e.g. the computational complexity increase in large models) of character-based tokenization, we of course do not want to replace sub-word tokenization. However, we believe that our models deserve a place in the toolbox of developmentally more plausible language models. They can be used to test what kind of linguistic knowledge

[7] Our phoneme data does not include word stress.

can be learned from raw input and answer questions about the learnability of linguistic knowledge from an even poorer stimulus (Thomas, 2002; Berwick et al., 2011) than the "stimulus" of subword models. In combination with phoneme representations, they open up new avenues of inquiry, e.g. for phenomena on the phonological/phonetic or lexical levels of linguistic analysis – phenomena which are not captured by the coarse-grained structure of subword tokens. Moreover, character-based language models open new pathways into experiments with multilingual models. The Latin script, for example, offers a shared vocabulary for many languages, whereas the IPA even offers a shared vocabulary for practically all languages.

Limitations

As previously mentioned, our results are only snapshots of individual training runs. Repeated training efforts with different initialization would be needed to filter noise from actual tendencies.

Besides, in the light of the current BabyLM challenge, we could only test these phenomena for English. The differences between grapheme and phoneme models may not generalize to other languages with different writing systems, languages with different levels of phonemic correspondences and systematicity in their orthography (like English or French vs Spanish or Czech), and languages with different morpho-phonological systems.

Acknowledgements

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

References

Manex Agirrezabal, Sidsel Boldsen, and Nora Hollenstein. 2023. The Hidden Folk: Linguistic Properties Encoded in Multilingual Contextual Character Representations. In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, pages 6–13, Toronto, Canada. Association for Computational Linguistics.

Lucas F.E. Ashby, Travis M. Bartley, Simon Clematide, Luca Del Signore, Cameron Gibson, Kyle Gorman, Yeonju Lee-Sikka, Peter Makarov, Aidan Malanoski, Sean Miller, Omar Ortiz, Reuben Raff, Arundhati Sengupta, Bora Seo, Yulia Spektor, and Winnie Yan. 2021. Results of the second SIGMORPHON shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 115–125, Online. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Robert C. Berwick, Paul Pietroski, Beracah Yankama, and Noam Chomsky. 2011. Poverty of the Stimulus Revisited. *Cognitive Science*, 35(7):1207–1242.

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 7–18, Singapore. Association for Computational Linguistics.

Lou Burnard. 2007. Reference Guide for the British National Corpus (XML Edition).

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every Layer Counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 210–224, Singapore. Association for Computational Linguistics.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [Call for Papers] The 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Preprint*, arXiv:2404.06214.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an Efficient Tokenization-Free Encoder for Language Representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

David Crystal. 2015. *Making a Point: The Persnickety Story of English Punctuation*, 1 edition. St. Martin's Press, New York.

Lukas Edman and Lisa Bylinina. 2023. Too Much Information: Keeping Training Simple for BabyLMs. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 61–69, Singapore. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6903–6915, Barcelona, Spain (Online).

International Committee on Computational Linguistics.

Ralph H. Emerson. 1997. English spelling and its relation to sound. *American Speech*, 72(3):260–288.

Clayton Fields, Osama Natouf, Andrew McMains, Catherine Henry, and Casey Kennington. 2023. Tiny Language Models Enriched with Multimodal Knowledge from Multiplex Networks. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 19–29, Singapore. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Martin Gerlach and Francesc Font-Clos. 2020. A Standardized Project Gutenberg Corpus for Statistical Analysis of Natural Language and Quantitative Linguistics. *Entropy*, 22(1):126.

Kyle Gorman, Lucas F.E. Ashby, Aaron Goyzueta, Arya McCarthy, Shijie Wu, and Daniel You. 2020. The SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 40–50, Online. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

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.

International Phonetic Association, editor. 1999. *The Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. Elements of World Knowledge (EWOK): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint*.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Arya D. McCarthy, Jackson L. Lee, Alexandra DeLucia, Travis Bartley, Milind Agarwal, Lucas F.E. Ashby, Luca Del Signore, Cameron Gibson, Reuben Raff, and Winston Wu. 2023. The SIGMORPHON 2022 shared task on cross-lingual and low-resource grapheme-to-phoneme conversion. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 230–238, Toronto, Canada. Association for Computational Linguistics.

Roger K. Moore and Lucy Skidmore. 2019. On the use/misuse of the term 'phoneme'. In *Interspeech 2019*, pages 2340–2344. ISCA.

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.

Linh The Nguyen, Thinh Pham, and Dat Quoc Nguyen. 2023. XPhoneBERT: A Pre-trained Multilingual Model for Phoneme Representations for Text-to-Speech. *Preprint*, arXiv:2305.19709.

Aidan Pine, Patrick William Littell, Eric Joanis, David Huggins-Daines, Christopher Cox, Fineen Davis, Eddie Antonio Santos, Shankhalika Srikanth, Delasie Torkornoo, and Sabrina Yu. 2022. Gi22Pi Rule-based, index-preserving grapheme-to-phoneme transformations Rule-based, index-preserving grapheme-to-phoneme transformations. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–60, Dublin, Ireland. Association for Computational Linguistics.

Ernst Pulgram. 1951. Phoneme and grapheme: A parallel. *WORD*, 7(1):15–20.

Iggy Roca. 2016. Phonology and English spelling. In Vivian Cook and Des Ryan, editors, *The Routledge Handbook of the English Writing System*, 1 edition, pages 65–91. Routledge, London.

Phillip Rust, Jonas F. Lotz, Emanuele Bugliarello, Elizabeth Salesky, Miryam de Lhoneux, and Desmond Elliott. 2023. Language modelling with pixels. In *The Eleventh International Conference on Learning Representations*.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, 26(3):339–373.

Mukuntha Narayanan Sundararaman, Ayush Kumar, and Jithendra Vepa. 2021. PhonemeBERT: Joint Language Modelling of Phoneme Sequence and ASR Transcript. In *Interspeech 2021*, pages 3236–3240. ISCA.

Yi Tay, Vinh Q. Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2022. Charformer: Fast character transformers via gradient-based subword tokenization. In *International Conference on Learning Representations*.

Margaret Thomas. 2002. Development of the concept of "the poverty of the stimulus". *The Linguistic Review*, 19(1-2):51–71.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby Llama: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 251–261, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

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.

Richard L. Venezky. 1967. English orthography: Its graphical structure and its relation to sound. *Reading Research Quarterly*, 2(3):75–105.

Ömer Veysel Çağatan. 2023. ToddlerBERTa: Exploiting BabyBERTa for Grammar Learning and Language Understanding. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 143–151, Singapore. Association for Computational Linguistics.

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.

Hongyu Wang, Shuming Ma, Li Dong, Shaohan Huang, Huaijie Wang, Lingxiao Ma, Fan Yang, Ruiping Wang, Yi Wu, and Furu Wei. 2023. BitNet: Scaling 1-bit Transformers for Large Language Models. *Preprint*, arXiv:2310.11453.

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.

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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. 2022. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech. In *Interspeech 2022*, pages 456–460. ISCA.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. In *Interspeech 2022*, pages 446–450. ISCA.

A Full BLiMP scores

Phenomenon	Graph. model	Graph. model, no whitesp.	Phon. model	Phon. model, no whitesp.
BLiMP	**71.69%**	68.88%	66.90%	64.88%
BLiMP supplement	52.30%	**56.28%**	55.42%	54.13%
adjunct_island_filtered	73.17%	**76.72%**	35.24%	36.75%
anaphor_gender_agreement_filtered	85.48%	82.29%	**86.30%**	69.10%
anaphor_number_agreement_filtered	**97.10%**	88.51%	95.17%	87.00%
animate_subject_passive_filtered	68.60%	**71.62%**	68.83%	62.91%
animate_subject_trans_filtered	**91.01%**	90.57%	82.23%	77.79%
causative_filtered	**69.07%**	68.09%	66.01%	64.55%
complex_NP_island_filtered	43.38%	**47.28%**	38.30%	43.85%
coordinate_structure_constraint_complex_left_branch_filtered	**46.36%**	37.75%	36.31%	30.68%
coordinate_structure_constraint_object_extraction_filtered	62.38%	65.12%	**65.86%**	63.22%
determiner_noun_agreement_1_filtered	97.31%	**97.74%**	52.85%	52.85%
determiner_noun_agreement_2_filtered	96.99%	**97.10%**	85.61%	82.81%
determiner_noun_agreement_irregular_1_filtered	**83.85%**	78.12%	72.25%	70.78%
determiner_noun_agreement_irregular_2_filtered	**90.00%**	87.56%	84.15%	76.59%
determiner_noun_agreement_with_adj_2_filtered	**92.24%**	90.75%	79.81%	76.94%
determiner_noun_agreement_with_adj_irregular_1_filtered	**82.45%**	77.30%	73.96%	71.17%
determiner_noun_agreement_with_adj_irregular_2_filtered	**82.38%**	78.93%	72.26%	69.88%
determiner_noun_agreement_with_adjective_1_filtered	**94.96%**	91.00%	51.77%	51.55%
distractor_agreement_relational_noun_filtered	**86.29%**	45.05%	68.40%	57.11%
distractor_agreement_relative_clause_filtered	**58.09%**	43.17%	50.98%	57.41%
drop_argument_filtered	75.76%	**75.98%**	60.87%	62.07%
ellipsis_n_bar_1_filtered	51.50%	**56.36%**	54.36%	53.87%
ellipsis_n_bar_2_filtered	58.09%	**63.29%**	43.36%	49.64%
existential_there_object_raising_filtered	**81.65%**	72.66%	79.80%	68.10%
existential_there_quantifiers_1_filtered	**99.46%**	97.42%	96.77%	93.76%
existential_there_quantifiers_2_filtered	28.21%	33.92%	38.42%	**43.69%**
existential_there_subject_raising_filtered	83.98%	82.90%	**84.31%**	80.84%
expletive_it_object_raising_filtered	70.09%	**73.12%**	72.46%	70.22%
inchoative_filtered	**55.79%**	52.28%	44.91%	46.67%
intransitive_filtered	**68.32%**	67.17%	46.31%	50.58%
irregular_past_participle_adjectives_filtered	**94.80%**	88.14%	72.84%	63.58%
irregular_past_participle_verbs_filtered	81.53%	81.10%	**85.14%**	77.39%
irregular_plural_subject_verb_agreement_1_filtered	**83.33%**	76.62%	82.21%	72.14%
irregular_plural_subject_verb_agreement_2_filtered	**89.46%**	87.33%	88.00%	83.86%
left_branch_island_echo_question_filtered	65.15%	61.67%	63.15%	**70.86%**
left_branch_island_simple_question_filtered	**60.15%**	46.79%	57.83%	50.26%
matrix_question_npi_licensor_present_filtered	15.82%	12.38%	17.98%	**31.75%**
npi_present_1_filtered	**50.39%**	40.59%	46.75%	48.51%
npi_present_2_filtered	49.89%	**50.33%**	45.62%	48.69%
only_npi_licensor_present_filtered	**98.07%**	48.64%	76.87%	92.06%
only_npi_scope_filtered	50.90%	44.92%	61.05%	**80.53%**
passive_1_filtered	89.17%	**90.60%**	87.74%	86.79%
passive_2_filtered	88.15%	**89.37%**	83.61%	81.28%
principle_A_c_command_filtered	55.07%	**59.51%**	51.48%	59.41%
principle_A_case_1_filtered	**100.00%**	**100.00%**	**100.00%**	99.89%
principle_A_case_2_filtered	91.58%	**92.57%**	88.20%	78.80%
principle_A_domain_1_filtered	96.39%	98.25%	**100.00%**	**100.00%**
principle_A_domain_2_filtered	53.55%	50.71%	**63.61%**	51.80%
principle_A_domain_3_filtered	50.90%	50.90%	**61.00%**	55.58%
principle_A_reconstruction_filtered	41.88%	34.64%	**53.67%**	47.67%
regular_plural_subject_verb_agreement_1_filtered	**93.48%**	90.45%	88.76%	80.11%
regular_plural_subject_verb_agreement_2_filtered	**90.37%**	85.19%	82.65%	77.67%
sentential_negation_npi_licensor_present_filtered	96.19%	96.74%	**99.35%**	96.52%
sentential_negation_npi_scope_filtered	21.70%	23.08%	33.30%	**40.76%**
sentential_subject_island_filtered	40.89%	39.33%	**58.17%**	57.54%
superlative_quantifiers_1_filtered	66.70%	66.80%	**70.99%**	54.14%
superlative_quantifiers_2_filtered	76.37%	**83.77%**	69.98%	61.16%
tough_vs_raising_1_filtered	**36.50%**	28.80%	23.73%	29.32%
tough_vs_raising_2_filtered	81.41%	**82.93%**	80.76%	78.37%
transitive_filtered	**80.07%**	74.77%	70.85%	66.94%
wh_island_filtered	61.77%	**63.54%**	61.04%	38.75%
wh_questions_object_gap_filtered	78.70%	75.20%	**80.33%**	76.37%
wh_questions_subject_gap_filtered	92.32%	**92.54%**	92.43%	90.31%
wh_questions_subject_gap_long_distance_filtered	91.60%	93.35%	93.58%	**94.87%**
wh_vs_that_no_gap_filtered	95.82%	95.93%	**96.17%**	94.54%
wh_vs_that_no_gap_long_distance_filtered	94.86%	**97.37%**	96.57%	94.74%
wh_vs_that_with_gap_filtered	**27.20%**	26.01%	5.55%	7.07%
wh_vs_that_with_gap_long_distance_filtered	**7.03%**	4.18%	3.41%	4.62%
supplement_hypernym	51.19%	**51.90%**	51.07%	51.19%
supplement_qa_congruence_easy	48.44%	54.69%	56.25%	**57.81%**
supplement_qa_congruence_tricky	26.67%	**39.39%**	25.45%	25.45%
supplement_subject_aux_inversion	78.54%	77.22%	**86.11%**	79.75%
supplement_turn_taking	56.79%	**58.21%**	**58.21%**	56.43%