What Should Baby Models Read? Exploring Sample-Efficient Data Composition on Model Performance

Hong Meng Yam
Stanford University
hongmeng@stanford.edu

Nathan Paek
Stanford University
nathanjp@stanford.edu

Abstract

We explore the impact of pre-training data composition on the performance of small language models in a sample-efficient setting. Using datasets limited to 10 million words, we evaluate several dataset sources—including child-directed speech (CHILDES), classic books (Gutenberg), synthetic data (TinyStories), and a mix of these (Mix)—across different model sizes ranging from 18 million to 705 million parameters. Our experiments show that smaller models (e.g., GPT2-18M and GPT2-44M) benefit from training on diverse datasets like Mix, achieving better performance on linguistic benchmarks. In contrast, larger models (e.g., GPT2-97M, GPT2-705M, and LLaMA-360M) perform better when trained on more complex and rich datasets like Gutenberg. Models trained on the CHILDES and TinyStories datasets underperformed across all model sizes. These findings suggest that the optimal dataset for sample efficient training depends on the model size, and that neither child-directed speech nor simplified stories are optimal for language models of all sizes. We highlight the importance of considering both dataset composition and model capacity for effective sample efficient language model training.

1 Introduction

In recent years, advancements in natural language processing have been largely driven by scaling language models to unprecedented sizes. Various large-language model (LLM) scaling laws have been formulated (Sardana et al., 2024), with perhaps the most influential being the Chinchilla law, which demonstrates that parameters and tokens scale approximately linearly as the model scales (Hoffmann et al., 2024). Many subsequent LLMs have been trained following this model (Rae et al., 2021), with some models including the Llama 2 and Llama 3 family of models being trained on 2 and 15 trillion tokens respectively, far more than the 'optimal' amount according to the Chinchilla scaling law (Dubey et al., 2024). However, it is often prohibitive to train such large models, and impractical to continue scaling with the amounts of data required to train such models.

This has sparked interest in small language models (Schick and Schütze, 2021; Magister et al., 2023) with much fewer parameters, requiring much less data for training. While much research has been conducted on knowledge distillation and improving the model architecture for small language models, comparably less research has investigated the contributions of different types of data used for model training, which is arguably just as important. Indeed, because LLM pretraining data typically comprises a mix of sources (Chowdhery et al., 2023), researchers have found that the composition of pretrained data greatly affects model performance (Du et al., 2022; Wei et al., 2015), though determining the optimal recipe for pretraining data is challenging. Recent research exploring optimization of pretraining data for LLMs at scale includes DoReMi, which trains a small proxy model to produce domain weights for downstream tasks, and then uses the model to resample the dataset for training huge LLMs (Xie et al., 2024). However, the question of how to choose data for sample-efficient training of small language models, such as in cases where computational resources are limited, has received little attention.

Psycholinguistic precedent exists for sample-efficient pretraining; children see much less words than a modern LLM yet perform exceptionally well on reasoning tasks. For example, Chinchilla sees over 10000 times the number of words a 13 year old child has ever encountered (Choshen et al., 2024). By the time typical English-speaking children at around 6 years old have obtained adult-level grammatical knowledge (Kemp et al., 2005), they have seen only around 10-50M words (Hart et al., 1997; Huebner et al., 2021). In comparison, Llama-3 is

The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, pages 284–291

trained on 15T tokens (Dubey et al., 2024). Given the great disparity between the amount of training data an LLM requires and what children require, it seems worthwhile to investigate whether training LLMs can be as sample efficient.

BabyBERTa (Huebner et al., 2021) attempts to address this, showing that when training a model on data similar to what is seen by children between the ages 1 and 6, it is able to acquire grammatical knowledge similar to pretrained RoBERTa-base, but with around 15X fewer parameters and 6,000X fewer words; this indicates that utilizing child-directed input may be advantageous for more sample efficient pretraining (Huebner et al., 2021). Similarly, Eldan and Li (2023) follow suit, releasing TinyStories, a synthetic dataset of short stories that only contain words that typical 3- to 4-year-old children understand. They demonstrate that TinyStories can be leveraged to train language models with much less parameters than SOTA models, yet still produce coherent output with almost perfect grammar as well as emergent reasoning abilities. Along the same vein, GPT-wee (Bunzeck and Zarrieß, 2023) shows that child-directed speech can be used with curriculum learning for simulating children's learning as a potential solution to sample-constrained training.

In this paper, we evaluate the effect of different datasets on model performance for sample efficient model training. In our case, we limit our training dataset to 10M words, in accordance with the BabyLM Challenge's super-strict track (Choshen et al., 2024). We consider several different types of datasets, namely child-directed speech (CHILDES), classic books (Gutenberg), a mixed dataset (Mix) and the TinyStories dataset. Experimental results show that smaller models benefit from training on diverse datasets like Mix on the BabyLM evaluation suite (Choshen et al., 2024), but larger models perform better when trained on more complex and rich datasets like Gutenberg. Our findings suggest that the optimal dataset depends on the model size and that neither child-directed speech nor child-directed stories are optimal for language models of any sizes.

2 Dataset

For our experiments, we obtained datasets from the BabyLM Challenge (Choshen et al., 2024). Individual categories of 10M-word datasets were procured by extracting the first 10M words from that category in the 100M-word dataset of the BabyLM challenge. We also used Mix, the 10M-word developmentally-plausible corpus of BabyLM, and TinyStories. To measure for complexity in the language of these datasets, we use several readability metrics, including the Flesch reading ease (FRE) score (Flesch, 1948), ARI (Automated Readability Index) (Smith and Senter, 1967), and the Gunning fog index (Gunning, 1969).

For a document $d_i \in \mathcal{C}$, its FRE score is computed as:

$$\text{FRE}(d_i) = 206.835 - (1.015 \cdot \text{ASL}) - (84.6 \cdot \text{ASW})$$

where ASL is the average sentence length (the number of words divided by the number of sentences) and ASW is the average number of syllables per word (the number of syllables divided by the number of words). Higher FRE scores correspond to simpler texts (e.g., children's literature), while lower scores indicate more complex writing (e.g., machine learning papers). The ARI score is calculated as:

$$\text{ARI}(d_i) = 4.71 \cdot \left(\frac{\text{characters}}{\text{words}}\right) + 0.5 \cdot \left(\frac{\text{words}}{\text{sentences}}\right) - 21.43$$

Higher ARI scores indicate more complex text requiring higher grade levels to comprehend. The Gunning fog index score is calculated as:

$$\text{Fog}(d_i) = 0.4 \cdot \left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100 \cdot \left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

Like ARI, higher Gunning fog scores indicate more complex text.

Our individual datasets comprise:

- **CHILDES**: The CHILDES dataset is composed of examples of the human language acquisition process starting from a very young age (MacWhinney, 2000). We constructed a 10 million word training corpus from the CHILDES portion of the small track (100M). We took the first 10M words from the CHILDES portion.
- **Gutenberg**: The Gutenberg dataset is a large dataset composed of English language books (Gerlach and Font-Clos, 2020). We took the first 10M words from the Gutenberg portion of the small track dataset.
- **Mix (Default)**: This was the default 10M dataset for the strict-small track. The split of is displayed below:

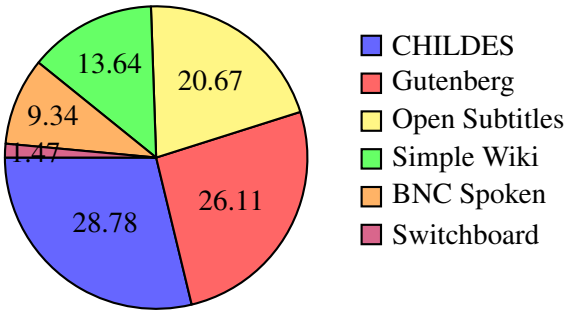

Figure 1: Default dataset composition

- **TinyStories**: We took the first 10M words from the TinyStories dataset on Hugging Face[1] (Eldan and Li, 2023). (FRE = 105.19)

Dataset	FRE	Gunning Fog	ARI
Mix	105.89	5.62	1.59
CHILDES	115.70	2.84	0.20
Gutenberg	87.49	9.89	7.12
TinyStories	105.19	4.83	0.85

Table 1: Readability metrics across different datasets. Lower FRE and higher Gunning Fog and ARI scores indicate a more complex dataset.

3 Methodology

3.1 Preprocessing

For both pre-processing and model training, we built off the BabyLlama repository[2] (Timiryasov and Tastet, 2023). Following their pre-processing steps, we applied regex-based cleaning and trained a Byte-Pair Encoding tokenizer on the training sets of whatever dataset we were working with. The train and dev sets were split into 128-token chunks, with the model being presented a new random permutation of these chunks in each epoch. Validation loss is computed at the end of each epoch using a fixed, randomly sampled subset of the dev set.

3.2 Training

Given that this builds upon the TinyStories paper, focused on dataset optimization for very small language models, we focused mainly on GPT models of sizes 18M, 44M and 97M, which we trained on various datasets. We used this to explore whether different model sizes would affect which dataset performed the best. We trained for 4 epochs, using consistent hyper-parameters. Subsequently, we trained a Llama-20M model to confirm that the same pattern regarding dataset complexity is observed in Llama models as well. Lastly, large model baselines of GPT2-705M and Llama-360M are used, as these were the original parent model sizes originally used by last year's BabyLM winning model (Timiryasov and Tastet, 2023).

3.3 Evaluation

Evaluation of model performance was done using the BabyLM evaluation suite (Choshen et al., 2024). This consists of the following benchmarks:

- **BLiMP:** BLiMP (Benchmark of Linguistic Minimal Pairs for English) evaluates language models on their ability to identify grammatical acceptability. It presents pairs of sentences that differ by one linguistic element, testing the model's understanding of 12 areas of English morphology, syntax, and semantics, such as anaphor agreement and filler-gap constructions. It measures how well models assign higher probability to the grammatically correct sentence in each pair. (Warstadt et al., 2020)

- **EWoK:** EWoK (Elements of World Knowledge) evaluates language models on their ability to build and apply internal world models. It tests models' understanding of concepts and contexts by presenting them with minimal pairs of scenarios where the models determine the plausibility of context and target combinations. (Ivanova et al., 2024)

- **GLUE:** GLUE (General Language Understanding Evaluation) evaluates language models on a variety of natural language understanding tasks. It covers tasks such as sentiment analysis, text similarity, question answering, and textual entailment. (Wang et al., 2018) Unlike in the BabyLM evaluation suite, however, we do not do finetuning in this case and run it as a zero-shot evaluation due to computational constraints.

4 Results and Discussion

Overall, our results demonstrate that the effectiveness of a training dataset is dependent on the model size. Specifically, smaller models (with fewer parameters) benefit more from training on a diverse

[1] https://huggingface.co/datasets/roneneldan/TinyStories

[2] https://github.com/timinar/BabyLlama

Model	Dataset	BLiMP Supplement	BLiMP Filtered	EWoK	Macroaverage
GPT2-18M	CHILDES	52.8	58.2	50.5	53.83
	Gutenberg	55.7	62.4	50.3	56.13
	Mix	**55.9**	**63.7**	49.7	**56.43**
	TinyStories	55.2	57.5	**50.7**	54.47
GPT2-44M	CHILDES	55.3	57.8	**51.2**	54.77
	Gutenberg	57.6	63.0	50.0	56.87
	Mix	**58.2**	**65.6**	50.4	**58.07**
	TinyStories	52.8	57.1	50.4	53.43
GPT2-97M	CHILDES	49.7	60.5	49.6	53.27
	Gutenberg	**59.0**	**65.3**	**51.1**	**58.47**
	Mix	58.0	66.0	50.6	58.20
	TinyStories	54.6	59.1	50.3	54.67
Llama-20M	CHILDES	53.4	57.9	50.2	53.83
	Gutenberg	**57.4**	60.0	**50.6**	**56.00**
	Mix	56.6	**62.8**	50.2	56.53
	TinyStories	46.7	51.1	49.8	49.20
GPT2-705M	Gutenberg	**59.9**	**66.8**	**50.6**	**59.10**
	Mix	56.7	66.1	50.6	57.80
Llama-360M	Gutenberg	**56.7**	**66.5**	50.2	**57.80**
	Mix	56.6	62.8	**50.5**	56.63

Table 2: Summary of BLiMP filtered, BLiMP supplement, EWoK results, and Macroaverage for various models and datasets

dataset like Mix, while larger models show improved performance when trained on the Gutenberg dataset. As shown in Table 2, for smaller models like GPT2-18M and GPT2-44M, Mix consistently achieves the best performance on BLiMP, scoring 63.7 and 65.6 respectively on BLiMP Filtered, and 55.9 and 58.2 on BLiMP Supplement. However, as we move to larger models like GPT2-97M and GPT2-705M, the Gutenberg dataset takes the lead, achieving the highest scores across most metrics (59.0 and 59.9 on BLiMP Supplement, 65.3 and 66.8 on BLiMP Filtered). We see this also extend to the Llama models as well, where the larger Llama-360M performs best with Gutenberg data (56.7 on BLiMP Supplement and 66.5 on BLiMP Filtered), while the smaller Llama-20M shows mixed results between Gutenberg and Mix. Interesting, both CHILDES and TinyStories consistently underperform across all model sizes, with scores typically lower than both Mix and Gutenberg datasets. On the other hand, we see a very different story when looking at macro average GLUE scores for the models (Table 3), with TinyStories performing well for small models and CHILDES performing well for the big model. However, when examining the GLUE subtasks further, we do not see a clear trend on which dataset type results a stronger performance, and cannot conclude a clear trend here.

4.1 Dataset and model performance

Model performance results on various datasets was observed in table 2. Small models, such as GPT2-18M and GPT2-44M, have limited capacity due to fewer parameters. This constraint affects their ability to capture complex linguistic patterns and nuanced language structures. Datasets like Gutenberg with a relatively lower FRE score (87.49) contain wider vocabulary, more intricate syntax, and nuanced semantic meaning. Due to their limited capacity, small models cannot fully learn from the complexity of the dataset. They oversimplify the language patterns, leading to high bias and poor generalization. This underfitting results in lower performance on evaluation benchmarks.

In contrast, larger models, such as GPT2-97M, GPT2-705M, and LLaMA-360M, possess greater capacity to learn and represent complex patterns due to their increased number of parameters. Because the Gutenberg dataset, consisting of a diver-

Model	Dataset	MRPC	MultiRC	QNLI	SST-2	BoolQ	MNLI	QQP	WSC	RTE	Cola (MCC)	Macro Average
GPT2-18M	CHILDES	34.31	45.50	50.92	53.67	58.59	32.42	42.47	38.46	48.20	-0.07	40.45
	Gutenberg	35.78	52.56	49.52	47.94	46.73	32.74	60.68	61.54	53.24	0.05	44.08
	Mix	58.33	44.35	47.14	47.71	57.00	32.42	46.77	46.15	44.60	0.03	42.45
	TinyStories	60.78	42.86	51.72	51.83	62.63	32.56	50.54	42.31	48.20	0.06	44.35
GPT2-44M	CHILDES	46.57	42.41	51.13	47.71	55.96	32.84	41.52	53.85	46.76	0.07	41.88
	Gutenberg	64.71	45.54	50.88	50.92	60.98	31.85	37.32	38.46	43.88	-0.02	42.45
	Mix	52.94	47.07	50.62	48.17	55.23	32.42	54.06	38.46	42.45	0.03	42.14
	TinyStories	45.59	53.09	47.04	48.39	42.26	33.19	62.01	59.62	54.68	-0.06	44.58
GPT2-97M	CHILDES	57.35	53.42	49.27	50.23	44.59	35.76	62.47	61.54	53.96	0.06	46.86
	Gutenberg	54.90	47.69	50.62	53.21	54.98	31.46	38.67	38.46	43.17	0.03	41.32
	Mix	47.05	49.88	48.57	50.00	44.10	33.48	61.82	61.54	56.12	-0.05	45.25
	TinyStories	65.20	43.61	50.40	51.83	62.08	32.03	38.05	44.23	50.36	0.07	43.79

Table 3: Detailed GLUE scores for various GPT models and datasets

sity of subject materials (Gerlach and Font-Clos, 2020), offers the most nuanced sentence structures and vocabulary out of all the datasets, it could be argued that diversity within the dataset may be more important than having a diverse basket of datasets for models with a higher number of parameters.

4.2 Dataset Convergence

In our experiments, CHILDES converged faster than either then Gutenberg or the Mix datasets for both GPT2-44M and GPT2-18M models. This can be observed in figure 2 and 3 below, and can be explained by the nature of CHILDES dataset. The higher FRE score (115.70) of this child-directed speech dataset indicates simpler grammatical structures, shorter sentences, and straightforward syntax compared to the adult-oriented language found in datasets like Gutenberg or Mix. In addition, because caregivers frequently repeat words and phrases when interacting with children, the dataset is characterized by high repetition, making the learning task of capturing the underlying structures and relationships in the data easier and faster to converge quickly during training. In short, due to the low perplexity of the CHILDES dataset, the model has less uncertainty in predicting the next word in a sequence, resulting in a smoother loss landscape and simplifying the learning task.

4.3 Underperformance of Child-directed and Synthetic Datasets

Neither the CHILDES nor TinyStories datasets performed very well on the BLiMP or EWoK evaluation suite (Choshen et al., 2024). The CHILDES dataset consistently underperformed no matter the model size, suggesting that child-directed speech may not be not advantageous for training a robust model. This is consistent with the lack of success

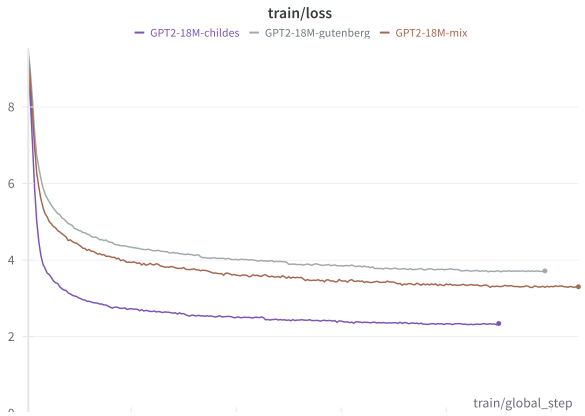

Figure 2: Train loss when training GPT2-18M on various datasets

in implementing curriculum learning for child data in the previous BabyLM challenge (Bunzeck and Zarrieß, 2023). In their paper, Bunzeck and Zarrieß noted that the integration of more sophisticated linguistic factors into the training process might be needed, as their curriculum approach based on prototypicality measures didn't effectively capture the language acquisition process they were looking for.

Considering the strong performance of TinyStories in (Eldan and Li, 2023), and the fact that we adopted the same GPT-44M architecture as in paper, with a hidden size of 768, 2 layers and 8 heads, we were surprised by the poor performance of the TinyStories dataset. That said, we only used a 10M subset of TinyStories, and given its limited vocabulary and grammatical range (and higher FRE score of 105.19), perhaps there was insufficient diversity and exposure to new formats as previously discussed. Additionally, we utilized different benchmarks. The BLiMP and EWoK benchmarks assess a model's understanding of complex grammatical rules and world knowledge; this is not likely to be adequately covered by the TinyStories dataset. In

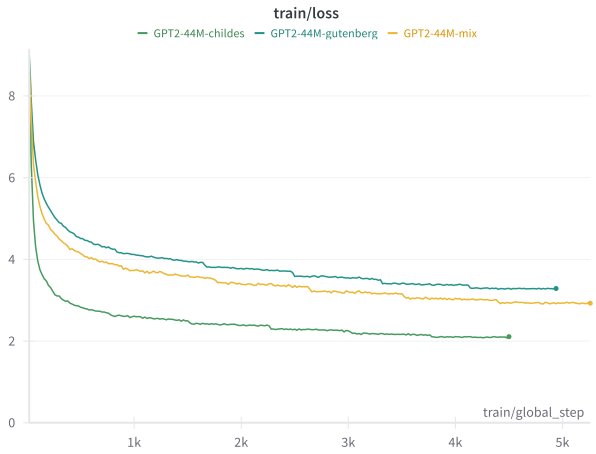

Figure 3: Train loss when training GPT2-44M on various datasets

short, models trained on TinyStories may lack exposure to the types of linguistic phenomena these benchmarks evaluate.

The disparity in TinyStories' performance across benchmarks likely stems from the divergent linguistic and cognitive demands of each dataset. GLUE evaluates general-purpose natural language understanding (NLU) tasks, such as sentiment analysis and paraphrase identification, which align well with the broad, semantic patterns learned from narrative content in TinyStories. In contrast, BLiMP emphasizes fine-grained syntactic and grammatical competence, while EWoK assesses factual reasoning and contextual world knowledge—skills that TinyStories' simplified narrative structure and limited syntactic diversity do not comprehensively support. Consequently, while TinyStories provides effective training for NLU, it lacks the complexity required for the precise linguistic and knowledge-based reasoning assessed by BLiMP and EWoK.

On the whole, however, we do not see the huge performance gains that were reported in the original TinyStories paper. The success of TinyStories in the original paper may perhaps be partially attributed to the narrative structure of the data, which provides contextual coherence and sequential dependencies that models can leverage. However, given that the Gutenberg dataset also contains narrative texts but with more complicated language and storylines, it offers better training data for models to learn general language patterns.

5 Limitations

Our study has several limitations. First, we used consistent hyper-parameters across all experiments for comparability, but this may not have been optimal for each model-dataset pair. Tuning hyperparameters individually could have yielded better performance.

Second, the BLiMP and EWoK benchmark assess linguistic competence on tasks on represented in datasets such as TinyStories or CHILDES, potentially biasing the evaluation. In short, there is a mismatch between the training data afforded by child datasets and the test set.

Lastly, due to computational limitations, models were trained for only four epochs. Longer training might have allowed models to better capture the nuances of the datasets.

6 Conclusion and Future Work

In this paper, we investigated the impact of dataset composition on the performance of small language models in a sample-efficient training regime. By training models of varying sizes on different datasets limited to 10 million words, we sought to identify which types of data are most beneficial for language acquisition in resource-constrained settings.

We found that tiny models (e.g., GPT2-18M and GPT2-44M) performed best when trained on the Mix dataset, which offers a diverse combination of language inputs, while slightly larger small language models achieved superior performance when trained on the Gutenberg dataset, leveraging its linguistic richness. In contrast, models trained on CHILDES or TinyStories underperformed regardless of size.

For future work, a more thorough investigation of other types of data sources such as news articles, scientific texts, and conversational data might better tease out the optimal dataset for model performance. Additionally, it might be useful to explore curriculum learning, which presumable models the developmental process of a language learning child.

Widening the benchmarks beyond GLUE and BLiMP tasks to coherent text generation, as well as scaling dataset sizes and tasks would allow for a more comprehensive and robust study as well.

Acknowledgments

We thank Stanford University for their support for this paper.

References

Bastian Bunzeck and Sina Zarrieß. 2023. Gpt-wee: How small can a small language model really get? In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 35–46.

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. *arXiv preprint arXiv:2404.06214*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Rudolf Franz Flesch. 1948. A new readability yardstick. *The Journal of applied psychology*, 32 3:221–33.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1).

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Betty Hart, Todd R Risley, and John R Kirby. 1997. Meaningful differences in the everyday experience of young american children. *Canadian Journal of Education*, 22(3):323.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2024. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

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.

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

Nenagh Kemp, Elena Lieven, and Michael Tomasello. 2005. Young children's knowledge of the" determiner" and" adjective" categories.

Brian MacWhinney. 2000. *The Childes Project: Tools for Analyzing Talk, Volume II: the Database*, 3rd edition. Psychology Press.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021.

Scaling language models: Methods, analysis & insights from training gopher. *ArXiv*, abs/2112.11446.

Nikhil Sardana, Jacob Portes, Sasha Doubov, and Jonathan Frankle. 2024. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws. *Preprint*, arXiv:2401.00448.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

E.A. Smith and R.J. Senter. 1967. *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

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. *Preprint*, arXiv:2308.02019.

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, 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.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. 2015. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pages 1954–1963. PMLR.

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