Call for Papers - The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus

https://babylm.github.io/

Alex Warstadt
ETH Zürich

Leshem Choshen
IBM Research

Aaron Mueller
Johns Hopkins University

Ethan Wilcox
ETH Zürich

Adina Williams
Meta AI

Chengxu Zhuang
MIT

Abstract

We present the call for papers for the BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. This shared task is intended for participants with an interest in small scale language modeling, human language acquisition, low-resource NLP, and cognitive modeling. In partnership with CoNLL and CMCL, we provide a platform for approaches to pretraining with a limited-size corpus sourced from data inspired by the input to children. The task has three tracks, two of which restrict the training data to pre-released datasets of 10M and 100M words and are dedicated to explorations of approaches such as architectural variations, self-supervised objectives, or curriculum learning. The final track only restricts the amount of text used, allowing innovation in the choice of the data, its domain, and even its modality (i.e., data from sources other than text is welcome). We will release a shared evaluation pipeline which scores models on a variety of benchmarks and tasks, including targeted syntactic evaluations and natural language understanding.

1 Motivation

Huge efforts have been put into optimizing LM pretraining at massive scales in the last several years (Raffel et al., 2020; Brown et al., 2020; Chowdhery et al., 2022; Hoffmann et al., 2022). While growing parameter counts often get the most attention, datasets have also grown by orders of magnitude. These increasingly larger pretraining datasets are visualized, to scale, in Figure 1. At the same time, there has been almost no progress in pretraining at smaller human-like data scales.

Focusing on scaled-down pretraining has several potential benefits: First, small-scale pretraining can be a sandbox for developing novel techniques that improve data efficiency. These techniques have the potential to then scale up to larger datasets commonly seen in applied NLP, and could be used
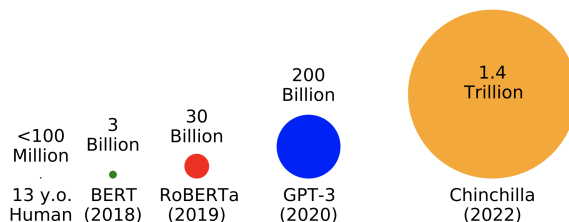

Figure 1: **Data Scale:** Modern Language Models are trained on data multiple orders of magnitude larger than the amount available to a typical human child. Image based off Fig. 1 from Warstadt and Bowman (2022)

to enhance current approaches to modeling low-resource languages. Second, improving our ability to train LMs on the same types and quantities of data that humans learn from will give us greater access to more plausible cognitive models of humans and help us understand what allows humans to acquire language so efficiently (Keller, 2010; Dupoux, 2018). That is, even model failure can help in developing hypotheses about the differences between human and LM language learning.

The goal of this shared task will be to incentivize researchers with an interest in pretraining and/or cognitive modeling to focus their efforts on optimizing pretraining given data limitations inspired by human development. Additionally, we hope to democratize research on pretraining—which is typically thought to be practical only for large industry groups—by drawing attention to open problems that can be addressed on a university budget.

2 Key Dates

- **January 2023:** Training data released
- **March 2023:** Evaluation pipeline released
- **July 15, 2023:** Results due
- **August 1, 2023:** Paper submissions due
- **Date TBA:** Presentation at CoNLL

Dataset	Domain	# Words		Proportion
		STRICT-SMALL	STRICT	
CHILDES (MacWhinney, 2000)	Child-directed speech	0.44M	4.21M	5%
British National Corpus (BNC),[1] dialogue portion	Dialogue	0.86M	8.16M	8%
Children's Book Test (Hill et al., 2016)	Children's books	0.57M	5.55M	6%
Children's Stories Text Corpus[2]	Children's books	0.34M	3.22M	3%
Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Written English	0.99M	9.46M	10%
OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	3.09M	31.28M	31%
QCRI Educational Domain Corpus (QED; Abdelali et al., 2014)	Educational video subtitles	1.04M	10.24M	11%
Wikipedia[3]	Wikipedia (English)	0.99M	10.08M	10%
Simple Wikipedia[4]	Wikipedia (Simple English)	1.52M	14.66M	15%
Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	0.12M	1.18M	1%
Total	–	9.96M	98.04M	100%

Table 1: The datasets we release for the STRICT and STRICT-SMALL tracks of the BabyLM Challenge. We present the number of words in the training set of each corpus that we include. [1]`http://www.natcorp.ox.ac.uk` [2]`https://www.kaggle.com/datasets/edenbd/children-stories-text-corpus` [3]`https://dumps.wikimedia.org/enwiki/20221220/` [4]`https://dumps.wikimedia.org/simplewiki/20221201/`

3 Tracks

This shared task includes three tracks: **STRICT**, **STRICT-SMALL**, and **LOOSE**.

The STRICT and STRICT-SMALL tracks require that submissions are trained exclusively on a fixed dataset, which we provide. The main difference between these tracks is the size of the dataset ($\sim$10M words vs. $\sim$100M words). Both datasets contain child-directed speech, transcribed speech from multiple sources, children's books, and Wikipedia, among other datasets. The STRICT-SMALL dataset is an approximately 10% uniform subsample of the STRICT dataset. See §4 for a full description of the fixed datasets. Winners will be determined based on performance on the shared evaluation set.

The LOOSE track relaxes these restrictions. Submissions must still be trained on a maximum of 100M words, and will be tested on the shared evaluation set. However, they are permitted to use unlimited non-linguistic data or text which differs from the restricted shared task. Training on additional text is allowed without limits if that text is generated by a model trained following the above restrictions. For this track, winners will be selected holistically based on evaluation performance, relevance to the shared task goals, potential impact, and novelty.

4 Dataset

We distribute a developmentally plausible pretraining dataset inspired by the input to children.[1] Submissions must use only this training data to be con-

sidered for the STRICT(-SMALL) tracks, but may use different data for the LOOSE track. The dataset has two key properties:

- **Under 100M words:** Children are exposed to 2M-7M words per year (Gilkerson et al., 2017). Choosing the beginning of adolescence (age 12) as a cutoff, the dataset should be between 24M-84M words.

- **Mostly transcribed speech:** Most of the input to children is spoken. Thus, we include a higher proportion of transcribed speech in our dataset.

The datasets we release are mixed domain, taken from multiple sources. Table 1 summarizes the composition of the datasets.

5 Evaluation

We will distribute a shared evaluation pipeline based in Google Colab. Colab provides access to relatively small GPUs; this will allow users from various research settings of varying resources to efficiently evaluate their submissions. Our evaluation code will also be public, such that those wishing to use their own computational resources may do so. More details about the evaluation pipeline and the set of tasks will be released subsequently.

The pipeline assumes all models can be loaded and queried in HuggingFace's `transformers` library (Wolf et al., 2020).[2] Additionally, all models must be able to score a sequence—e.g., assign

---

[1]Clicking on the following link will download the dataset (240MB zipped, 700MB unzipped): `https://github.com/babylm/babylm.github.io/raw/main/babylm_data.zip`

[2]While discouraged, participants whose models are not compatible with the `transformers` library can still conduct the necessary evaluation through their own pipeline.

a log-likelihood or pseudo log-likelihood (Wang and Cho, 2019; Salazar et al., 2020)—and must be able to be fine-tuned to perform classification tasks. Models do not need to be able to generate sequences. Submissions must include model outputs for each of the core evaluations in a format that we specify in our evaluation pipeline.

We choose evaluations that represent the core interests of this shared task, focusing on efficiency and applied NLP, as well as cognitive science, linguistics and language acquisition. Especially good performance in one but not both of these areas may be acknowledged with a special award.

5.1 Baselines

We will also release a series of baseline models with the evaluation pipeline. To train these, we simply take the hyperparameters from a series of established large language models and train them from scratch on our fixed datasets. We use hyperparameters from OPT (decoder-only; Zhang et al., 2022), RoBERTa (encoder-only; Liu et al., 2019), and T5 (encoder-decoder; Raffel et al., 2020). These are not meant to be strong baselines, but rather to provide a naïve starting point for improving language models for this domain.

6 Submissions

What you Need to Submit

- A link where we can download the model

- A .zip of predictions (from our eval pipeline)

- A short description of the approaches taken

- If LOOSE track: a link where we can download any additional data

---

Although scaled-down pretraining is more accessible to research groups with limited resources, pretraining is still expensive from a computational, energy, and financial perspective. To help groups plan for total costs, we will release an estimate of the resources required to pretrain on 10M words and 100M words. For the LOOSE track, evaluation of submissions may take into consideration computational efficiency as part of the holistic evaluation.

7 FAQs

Can papers be submitted to multiple tracks?
Yes. For example, a single paper can describe models which are submitted separately to the STRICT and STRICT-SMALL tracks.

Can I submit a paper about my work? Yes, we encourage all participants to submit their reports, which will be published in the proceedings of CoNLL. You may also describe any additional experiments beyond those required for the shared task evaluation.

Can I submit additional evaluation metrics? Yes, if you wish to submit your own evaluation metrics, along with model performance, alongside our standardized evaluation results these can be considered as part of the holistic evaluation in the LOOSE track.

What training regimes are permitted? For the STRICT/STRICT-SMALL tracks, any kind of training objective/regime is permitted, as long as the data restrictions are followed. Pretrained models may not be used for any purpose such as reranking or data augmentation.

We do however require for evaluation purposes that the model provides a function to score a sequence—e.g., log-likelihood for autoregressive models or pseudo-log-likelihood for masked language models—without the need for additional fine-tuning.

Are there any limits on hyperparameters? No. In the LOOSE track, parameter efficiency and training efficiency may be considered along with other factors in ranking submissions.

Are there any limits on the number of epochs? No. We put no restrictions on the number of epochs, for several reasons: First, from an engineering perspective, training LMs with SGD tends to require multiple epochs at these scales to achieve peak performance. Second, from a cognitive perspective, humans have a memory of linguistic experience, and can continue to access and learn from these memories. Third, we try not to make a stand on implementations to allow the most freedom for innovation.

8 Organizing Committee

Leshem Choshen   Aaron Mueller
Ryan Cotterell   Alex Warstadt
Kundan Krishna   Ethan Wilcox
Tal Linzen   Adina Williams
Haokun Liu   Chengxu Zhuang

Questions? Feel free to contact us at the following email addresses:

leshem.choshen@mail.huji.ac.il
haokunl@cs.unc.edu
amueller@jhu.edu
alexwarstadt@gmail.com
ewilcox@ethz.ch
chengxuz@mit.edu

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *Computing Research Repository*, arXiv:2204.02311.

Emmanuel Dupoux. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Computing Research Repository*, arXiv:1812.08092.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden. Association for Computational Linguistics.

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).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Brian MacWhinney. 2000. *The CHILDES project: The database*, volume 2. Psychology Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–371.

Alex Wang and Kyunghyun Cho. 2019. BERT has a mouth, and it must speak: BERT as a Markov random field language model. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 30–36, Minneapolis, Minnesota. Association for Computational Linguistics.

Alex Warstadt and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. 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.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pretrained transformer language models. *Computing Research Repository*, arXiv:2205.01068.