[Call for Papers] The 2ⁿᵈ BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus

https://babylm.github.io/

Leshem Choshen
IBM Research, MIT

Ryan Cotterell
ETH Zürich

Michael Y. Hu
NYU

Tal Linzen
NYU

Aaron Mueller
Northeastern University

Candace Ross
Meta AI

Alex Warstadt
ETH Zürich

Ethan Wilcox
ETH Zürich

Adina Williams
Meta AI

Chengxu Zhuang
MIT

Abstract

After last year's successful BabyLM Challenge, the competition will be hosted again in 2024/2025. The overarching goals of the challenge remain the same; however, some of the competition rules will be different. The big changes for this year's competition are as follows: First, we replace the loose track with a *paper track*, which allows (for example) non-model-based submissions, novel cognitively-inspired benchmarks, or analysis techniques. Second, we are relaxing the rules around pretraining data, and will now *allow participants to construct their own datasets* provided they stay within the 100M-word or 10M-word budget. Third, we introduce a *multimodal vision-and-language track*, and will release a corpus of 50% text-only and 50% image–text multimodal data as a starting point for LM model training. The purpose of this CfP is to provide rules for this year's challenge, explain these rule changes and their rationale in greater detail, give a timeline of this year's competition, and provide answers to frequently asked questions from last year's challenge.

1 The BabyLM Challenge

After last year's successful BabyLM Challenge, the competition will be hosted again in 2024/2025. The overarching goals of the challenge remain the same, however, some of the competition rules will be different. The purpose of this CFP is to provide a timeline of this year's competition, explain the rule changes and their rationale, and provide answers to frequently asked questions from last year's challenge.

As before, the goal of the BabyLM shared task is to incentivize researchers interested 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 pertaining by drawing attention to open problems that can be addressed on a university budget. Please see the previous call for paper (Warstadt et al., 2023a) or the introduction to last year's proceedings (Warstadt et al., 2023b) for a deeper discussion of the rationale behind the challenge.

This year's version of the BabyLM challenge will remain the same in terms of goals, but with a few key differences, which we list below:

- To encourage contributions that are related to the goals of the challenge, but do not involve direct competition entries, **we are introducing a paper-only track.** Paper track submissions could include things like novel cognitively-inspired evaluation metrics or in-depth analyses of one particular BabyLM model.

- Last year, all competition entrants were required to pre-train on a fixed corpus, meaning that submissions could not investigate the impact of data quality on pretraining. This year we will relax this requirement. While we will still provide language-only and multi-modal (see below) datasets of 100M and 10M words, **participants are free to construct their own datasets, provided that they stay within the 100M or 10M word budget.** Participants will be asked to provide a datasheet for any datasets they construct themselves.

- Human language learning is inherently multi-modal. To encourage more multi-modal submissions, **we are replacing last year's loose track with a vision-language track**. For this track, we will release a corpus of 50% text-only and 50% image–text multimodal data.

2 Key Dates

- **March 30 2024:** Training data released

- **April 30 2024:** Evaluation pipeline released

- **September 13 2024:** Results due

- **September 20 2024:** Papers due

- **October 8 2024:** Peer review begins

- **October 30 2024:** Peer review ends, acceptances and leaderboard released

- **December 9-15, 2024:** Presentations at NeurIPS (pending acceptance)

3 Participation Rules: Tracks

The second BabyLM Challenge includes three tracks: STRICT, STRICT-SMALL, and VISION and PAPER.

The STRICT and STRICT-SMALL tracks require that submissions be trained on 100M words or less (in STRICT) and 10M words or less (in STRICT-SMALL). These tracks no longer require that participants use the fixed dataset from last year's challenge, although we will still provide an updated version of this dataset for participants interested in using it. See Section 4 for details on the updated version of this language-only dataset. Models in these tracks will be evaluated on language-only evaluation tasks.

In the VISION track, participants will train multi-modal image-text models. Participants can use whatever training procedure they wish, as long as models can provide (pseudo) log-likelihoods to strings of text, conditioned on an image. Participants are free to use whatever data they wish, as long as the dataset is within a 100M word budget. To facilitate easier participation in this track, we will release a suggested multimodal dataset that consists of 50% text-only and 50% paired image–text data. Submissions to this track will be evaluated on language-only tasks, as well as multi-modal tasks.

In the PAPER track, participants are invited to submit a paper on a topic relevant to cognitive modeling with language models and scaled-down pertaining. Some submissions we have in mind include novel cognitively-inspired evaluation metrics or an in-depth analysis of one particular BabyLM model. Submissions to this track are welcome to include a BabyLM model, and to report its scores, however, evaluation task scores will not necessarily determine the best-paper awardee from this track.

4 Participation Rules: Provided Dataset

For this year, we update the dataset from the text-only dataset from the previous year and provide a novel multi-modal vision/text dataset for the VISION track. Note that, as long as participants stay under the 100M or 10M word budget, they are free to create their own datasets instead of using the ones provided.

For the text-only dataset updates, we replace the QED portion of the previous dataset with data from CHILDES.[1] This was done because, on further inspection, it was found that the QED dataset was of poorer quality than hoped for. CHILDES now comprises a larger portion of the new dataset, but the relative proportions of the rest of the data sources remain the same. Please see the previous year's CFP for more details about these data sources (Warstadt et al., 2023a)

In addition, we provide a novel image–text dataset to facilitate easier participation in the VISIONrack. This dataset includes two components: First, we provide 50M words of text-only data, which is downsampled from the 100M corpus (that is, we preserve the relative distribution from the different data sources). Second, we provide paired text–image data that includes 50M words of text. This dataset is drawn from two sources: 27M words from the Localized Narratives (Pont-Tuset et al., 2020b) and 23M words from the conceptual captions 3M dataset (Sharma et al., 2018b). For the Localized Narratives dataset, our challenge uses the text captions and the images from the MS-COCO (Lin et al., 2014) and Open Images (Kuznetsova et al., 2020) subset in this dataset. For the CC3M dataset, we used the image-caption pairs whose images were still valid in January 2024. We will provide the scripts to download the images. We will also provide visual embeddings of these images computed from a visual encoder pretrained using DINOv2 (Oquab et al., 2023), one of the SOTA unsupervised learning algorithms. In addition to the text-image data, we use half of the 100 million words from the text-only dataset as the text-only data in this track. We take the first half of every corpus within the text-only corpora mentioned above. Table 1 gives an overview of the datasets used.

If not using our pre-fixed datasets, participants will be required to submit a datasheet for their dataset (Gebru et al., 2021).

---

[1]Data can be downloaded from `https://osf.io/ad7qg/`

Dataset	Description	# Words (multimodal track)	# Words (strict track)	# Images
Localized Narratives(**Pont-Tuset et al., 2020a**)	Image Caption	27M	–	0.6M
Conceptual Captions 3M(**Sharma et al., 2018a**)	Image Caption	23M	–	2.3M
CHILDES (MacWhinney, 2000)	Child-directed speech	9M	17	–
British National Corpus (BNC), dialogue portion	Dialogue	4M	8M	–
Children's Book Test (Hill et al., 2016)	Children's books	3M	6M	–
Children's Stories Text Corpus	Children's books	2M	3M	–
Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2018)	Written English	5M	9M	–
OpenSubtitles (Lison and Tiedemann, 2016)	Movie subtitles	15M	31M	–
Wikipedia	Wikipedia (English)	5M	10M	–
Simple Wikipedia	Wikipedia (Simple English)	7M	15M	–
Switchboard Dialog Act Corpus (Stolcke et al., 2000)	Dialogue	<1M	1M	–
Total	–	100M	100M	2.9M

Table 1: Datasets for the multimodal and strict tracks of the 2nd BabyLM Challenge. Word counts are approximate and subject to slight changes.

5 Participation Rules: Evaluation

As in the previous year, we will distribute a shared evaluation pipeline based on Google Colab. For the STRICT and STRICT-SMALL track, evaluation tasks will be, largely, the same as the previous year. For the VISION track, we will include an additional set of multimodal evaluation tasks. As before, our evaluation code will be public, so that those who wish to use their own computational resources instead of the Google Colab may do so. More details about the evaluation pipeline and the set of tasks will be released subsequently.

In response to comments from the previous year, this year's evaluation pipeline will use catwalk, which will make it easier to submit models that do not conform to HuggingFace transformers library (although it will still be easy to submit models if they *do* conform to HuggingFace standards). All models must be able to assign a (pseudo) log-likelihood to a string of text 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.

5.1 Baselines

We will release a series of baseline models. As opposed to last year's baselines, which were trained relatively naively, this year's baseline will be based on the winning submissions from last year. For the STRICT and STRICT-SMALL tracks, we will release the following baselines: GPT2 (decoder-only; Radford et al., 2019), LTG-Bert (encoder-only; Samuel et al., 2023), and Contextualizer (decoder-only; Xiao et al., 2023). For the VISION track, we will release the GIT (Wang et al., 2022) and Flamingo (Alayrac et al., 2022) baselines. These baselines are meant to encourage participants to innovate and improve beyond existing models.

6 The Submission Process

6.1 Results submission

As with last year, submissions will be made using the Dynabench platform, which is an online platform for dynamic data collection and model benchmarking.[2]

What you Need to Submit:

- A link where we can download the model (any file-hosting service will do)
- Model predictions for our evaluation pipeline
- A datasheet for your dataset and a download link (if not using ours)

6.2 Paper submission

Submissions will be made through OpenReview. All submissions will be full archival papers and can be up to eight pages in length. Formatting requirements for submissions will be announced at a later date, once the presentation venue has been finalized. As before, we do allow dual-submission, however, we do not allow dual publication.

6.3 Review & Publication

BabyLM will hold its own review process. The acceptance criteria are based on soundness: We plan only to reject submissions that make incorrect or unjustified claims. Other feedback will be directed at the improvement of submissions.

7 FAQs

Can papers be submitted to multiple tracks? Yes. For example, a single paper can describe mod-

---

[2]https://dynabench.org/

els that are submitted separately to the STRICT and VISION tracks.

Can I submit a paper about my work? Yes, we require that all submissions be accompanied by a paper, which can be up to eight pages in length (although it doesn't have to be). Papers will be published in an archival format. All papers can describe experiments and analyses beyond the scope of the competition, and this type of non-competition contribution is required for the PAPER track.

Can I submit additional evaluation metrics? Yes, you may submit additional evaluation metrics alongside a competition model in the STRICT, STRICT-SMALL, and VISION tracks. This type of contribution is especially encouraged in the PAPER track.

What training regimes are permitted? Any kind of training objective/regime is permitted, as long as the data restrictions are followed. If you use ancillary models, for example for reranking or data augmentation, then the training data for these models is counted towards your 100M word budget.

For evaluation purposes, we require that the model provides a function to score a sequence of words without the need for additional fine-tuning.

Are there any limits on hyperparameters? No. But please share at the end what you found so we can learn from your efforts.

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. Our internal results suggest, however, that under regular circumstances, over-training on more than a couple epochs give minor gains at most.

Can I use external tools? Yes, but if they learned on language their tokens are counted towards the 100M. That means one can train on the same text both a tokenizer, a parser, an LM etc. or on parts of the 100M, but the sum of all text seen by all training can not surpass the amount of text allowed. This raises the questions of synthetic data. It is allowed, under some restrictions. You may generate the 100M tokens in any legal way you like (yes, distilling or writing your own is fair, if you figure what text facilitates learning it is interesting regardless of how to gather such text), you may also train eventually on more than 100M words by augmentation, however that only works in a closed system, i.e., the augmentors training data counts toward the limit, so for example training two LMs on half of the words, and then have them generate more words and train a model on both the original data and the new one is legit (and it was not tested in the previous competition, so even the example itself is interesting).

I have different modalities that can help If it is not linguistic data, prove it, last year's submission did not gain from non-linguistic grounding, but we encourage such scientific questions. If it is linguistic in nature (e.g., audio), then the words should still count towards the overall number of learnt words.

8 Organizing Committee

(Alphabetical by last name) Leshem Choshen, Ryan Cotterell, Michael Hu, Tal Linzen, Aaron Muller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, Chengzu Zhuang. Feel free to contact members of the organizing committee at: leshem.choshen@mail.huji.ac.il, aa.mueller@northeastern.edu, alexwarstadt@gmail.com, chengxuz@mit.edu

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

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.

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

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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: The database*, volume 2. Psychology Press.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020a. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020b. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: Bert meets british national corpus. *arXiv preprint arXiv:2303.09859*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018a. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018b. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Alex Warstadt, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox, and Chengxu Zhuang. 2023a. Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023b. 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 1–34.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2023. Towards more human-like language models based on contextualizer pretraining strategy. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 317–326.