Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora

Alex Warstadt[1*] **Aaron Mueller**[2,3*] **Leshem Choshen**[4,5] **Ethan Wilcox**[1] **Chengxu Zhuang**[4]
Juan Ciro[6] **Rafael Mosquera**[6] **Bhargavi Paranjape**[8]
Adina Williams[6,7] **Tal Linzen**[9] **Ryan Cotterell**[1]
[1]ETH Zürich [2]Northeastern University [3]Technion [4]MIT
[5]IBM Research [6]MLCommons [7]Meta AI (FAIR)
[8]University of Washington [9]New York University
warstadt@inf.ethz.ch aa.mueller@northeastern.edu

Abstract

Children can acquire language from less than 100 million words of input. Large language models are far less data-efficient: they typically require 3 or 4 orders of magnitude more data and still do not perform as well as humans on many evaluations. These intensive resource demands limit the ability of researchers to train new models and use existing models as developmentally plausible cognitive models. The BabyLM Challenge is a communal effort in which participants compete to optimize language model training on a fixed data budget. Submissions are compared on various evaluation tasks targeting grammatical ability, downstream task performance, and generalization. Participants can submit to up to three tracks with progressively looser data restrictions. From over 30 submissions, we extract concrete recommendations on how best to train data-efficient language models, and on where future efforts should (and perhaps should not) focus. The winning submissions using the LTG-BERT architecture (Samuel et al., 2023) outperformed models trained on trillions of words. Other submissions achieved strong results through training on shorter input sequences or training a student model on a pretrained teacher. Curriculum learning attempts, which accounted for a large number of submissions, were largely unsuccessful, though some showed modest improvements.

1 Introduction

Although there have massive improvements in the effectiveness of neural language models in the last decade, humans are still the state of the art in language learning. To achieve impressive results, language models need to be trained on hundreds of times more language input than a typical human will be exposed to in an entire lifetime. The BabyLM Challenge is a shared task that invites members of the natural language processing, linguistics, and cognitive science communities to train language models in low-resource data settings, where the amount of linguistic input resembles the amount received by human language learners. In doing so, our motivations (Section 2) are to improve the relevance of language models as cognitive models of human language acquisition, find more effective and data-efficient training algorithms for language models, and democratize research on language model training by emphasizing research questions that can be addressed on a smaller training budget.

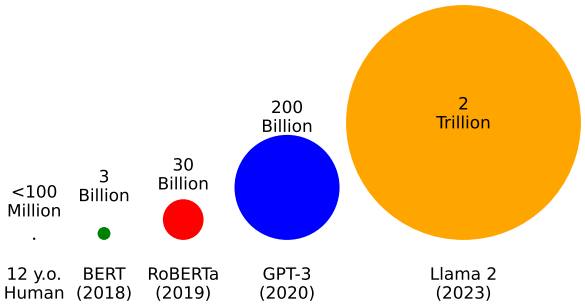

Figure 1: **Data Scale:** Modern Language Models are trained multiple orders of magnitude more word tokens than the amount available to a typical child. This image is based on Fig. 1 from Warstadt and Bowman (2022).

Participants in the shared task could submit to the *Strict*, *Strict-Small*, or *Loose* track, which, respectively, required models to be trained on corpora that constituted either 10 million words, 100 million words, or 100 million words plus an unlimited amount of additional non-linguistic data (Section 3). These corpora were constructed from a mixture of sources including developmentally plausible domains such as child-directed speech, transcribed dialogue, and children's literature (Section 4). To enable standardized evaluation and easy comparison of the resulting models, we create a leaderboard and release an evaluation pipeline (Section 5) targeting zero-shot grammatical performance, finetunability on language understanding

[*]Equal contribution.

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 1–34

tasks, and model inductive bias. We also contribute a novel set of zero-shot evaluation tasks targeting semantic and discourse-level phenomena.

We received 31 papers making a variety of contributions, ranging from designing novel architectures and tuning hyperparameters to employing curriculum learning and training teacher–student model pairs (Section 6). We conduct a meta-analysis of the results, yielding several concrete recommendations and scientific conclusions (Section 7). The winners of the challenge's various tracks made contributions that led to impressive improvements in our evaluation over not just the BabyLM baselines, but also the massively pretrained Llama 2 model (Touvron et al., 2023). The best-performing models overall (Charpentier and Samuel, 2023) use the LTG-BERT architecture (Samuel et al., 2023), which synthesizes a number of recent optimizations of the Transformer architecture. The winner of the *Loose* track (Xiao et al., 2023) trains the models continuously on the training samples belonging to the same source dataset while randomizing the dataset orders in each training epoch. Other submissions did not achieve strong downstream results, but still provided valuable scientific contributions. We received many curriculum learning submissions, including one that systematically tested a variety of strategies (Martinez et al., 2023) and reported few improvements over non-curriculum baselines. Steuer et al. (2023) found that benchmark performance is not correlated with a greater ability to predict human psycholinguistic data.

We plan to organize future BabyLM Challenges that will build on the success of this first iteration (Section 8). The winning submission from this year sets a high baseline for next year. Future iterations will need harder and more varied evaluations, including those that emphasize human-like processing and learning; they should emphasize new approaches that were not thoroughly explored this year, such as multimodality; and, they should incentivize compute-efficiency. Altogether, the first BabyLM Challenge has been a successful initiative, and we hope that this will continue to advance research on small-scale language models.

2 Motivation

The observation at the center of the BabyLM Challenge is this: Children are incredibly data-efficient language learners, and language models are not. Children are exposed to less than 100 million word tokens by age 13 (Gilkerson et al., 2017), while modern language models are typically trained on 3 or 4 orders-of-magnitude more data (Figure 1). This discrepancy raises two important questions: First, how is it that humans are able to learn language so efficiently? Second, what insights from human language learning can be used to improve language models?

A great deal of recent work in language model training seeks improvements by scaling up pretraining data and parameters (Raffel et al., 2020; Brown et al., 2020; Hoffmann et al., 2022; Touvron et al., 2023). Scaling is undoubtedly central to building deployable models (though see McKenzie et al. 2023 for counterexamples) and raises its own set of scientific questions, such as quantitative scaling laws (Kaplan et al., 2020) and the emergence of new abilities (Wei et al., 2022). However, increased emphasis on scaling is unlikely to lead to answers to the two questions we raised, and it excludes researchers without access to massive computational resources.

Thus, there are three principal benefits to data-limited language model training which the BabyLM Challenge aims to highlight:

1. Building more cognitively and developmentally plausible models of human language acquisition and processing,

2. Optimizing training pipelines prior to scaling by allowing for faster iteration on architectures and hyperparameters, and

3. Enabling research on language model training beyond highly funded industry groups.

Cognitive Modeling. Language models have been used to model aspects of human language learning and processing for decades (Elman, 1990; Hale, 2001; Reali and Christiansen, 2005, o.a.). While many researchers continue to advocate for language models as cognitive models (Keller, 2010; Dupoux, 2018; Linzen, 2019; Baroni, 2022; Warstadt and Bowman, 2022; Piantadosi, 2023; Wilcox et al., 2023), most agree that it is critical to make LMs learn in more human-like ways. Warstadt and Bowman (2022) and Linzen (2020) point to data quantity as the most egregious advantage that modern language models have over humans. When restricted to developmentally plausible data volumes, language models no longer perform well on benchmarks for human-like

syntactic and semantic behavior (van Schijndel et al., 2019; Zhang et al., 2021).

Working to close the data-efficiency gap between language models and humans will have two principal advantages for cognitive modeling. First, by reverse-engineering known and hypothetical aspects of the human learning scenario—from multimodal inputs and multi-agent interaction to innate linguistic structural biases—we can determine which factors are critical to our unique ability to learn language efficiently (Dupoux, 2018). Second, by minimizing differences between humans and models, we make results from controlled experiments carried out on models more likely to be applicable to humans (Warstadt and Bowman, 2022).

Faster iteration on architectures and hyperparameters for language modeling. Reducing the scale of training provides researchers with a sandbox in which to more fully explore this design space and better optimize training pipelines. The search space for design choices when training language models is enormous. Thus, it can be impractical, especially at large scales, to experiment with new model architectures, training objectives, or data preprocessing steps, in addition to necessary hyperparameter tuning. Models such as RoBERTa (Liu et al., 2019) have succeeded in making some optimizations to the BERT training pipeline, but more optimizations remain. Indeed, there are anecdotes of basic design choices for popular pipelines, such as the masking rate for BERT training (Wettig et al., 2023), being poorly tuned for years, despite hundreds or even thousands of papers using this training pipeline.

There are numerous dimensions along which to scale down training. Some works seek to optimize pipelines for a limited amount of compute, time, or money. Notable examples of such pipelines for bidirectional encoder-only include ELECTRA (Clark et al., 2020), 24-hour BERT (Izsak et al., 2021), and MosaicBERT (Portes et al., 2023). These pipelines typically combine multiple approaches, such as modifying training objectives to increase the number of supervised predictions per forward pass, using low-precision floating-point computations for certain components, reducing sequence length or padding, and altering the attention or feed-forward layers of the transformer block.

However, the objective of optimizing pipelines for a fixed data budget is relatively underexplored. This is changing in the last year with new models optimized for small datasets such as LTG-BERT (Samuel et al., 2023) and community-oriented events centered around data-limited training such as the Learning from Small Data workshop (Breitholtz et al., 2023) and the MiniPile Challenge (Kaddour, 2023).

Democratizing language model training research. The third goal of the BabyLM Challenge is to democratize research on pretraining—typically thought to be practical only for large industry groups—by drawing attention to challenging and important open problems that can be explored on a university budget. In recent years, efforts aimed at widening participation in LM research often take different avenues from the one proposed here, including aggregation of distributed computation power (Diskin et al., 2021), reliance on public computing infrastructure (Scao et al., 2022), aggregation of expertise, data and stepwise contributions (Don-Yehiya et al., 2023; Raffel, 2023) and modularity (Pfeiffer et al., 2023). Such a line of pretraining research proposes to keep costs large but to distribute them across funding sources through many contributing factors.

Other works on decentralizing computation (Diskin et al., 2021; Li et al., 2022; Lialin et al., 2023) or model recycling works generally take existing models and build upon them, proposing a single adaptation finetuning (Choshen et al., 2022), a single knowledge edit (De Cao et al., 2021), combining several models (Yadav et al., 2023), or iterative approaches showing that stacking such improvements can continually improve models (Don-Yehiya et al., 2023). Recently, a framework for doing so was also released (Kandpal et al., 2023). One can see the BabyLM challenge in this context as a suggestion to persist in using a centralized approach to pretraining, but making it tractable, by reducing the cost through increased focus on tractable research questions.

3 Guidelines and Timeline

Tracks. Submissions to BabyLM had to conform to one of three sets of guidelines, which we term **tracks**. In this section, we describe each competition track; for specific details about wording, see the original Call for Papers (Warstadt et al., 2023). The three tracks for the BabyLM challenges were ***Strict***, ***Strict-Small***, and ***Loose***. Participants in all tracks were allowed a constant number of English-language training tokens (100 million in *Strict* and

Loose and 10 million in *Strict-Small*) to be used in total for all software used in the pipeline. This data was released by the organizing committee and is described, in detail, in Section 4. *Loose* track submissions were encouraged to train on data beyond just the linguistic text data provided through the shared task (e.g., speech audio signal, code, music, or visual input). The *Loose* track also permitted the use of expert-annotated data, but any language data used to train the LM or auxiliary models counted towards the 100M word budget. Thus, for example, a *Loose* track submission could train a parser on the Penn Treebank (Marcus et al., 1993) and self-train to parse the pretraining corpus, as long as the number of words in the Penn Treebank plus the pretraining corpus total less than 100M.[1]

In general, seeing the same data twice (e.g., across different epochs) did not count as seeing more text. While it is unlikely that humans process data iteratively in a manner similar to epoch-based training, there is evidence that humans do repeat some of the information they process (e.g., in memory replay, Carr et al., 2011). Furthermore, epochs are very useful for gradient-based methods.

Finally, participants across all tracks were encouraged to submit models and papers even if their work did not fit into any of the three tracks. As the goal of the shared task is to advance efficient and cognitively plausible LM training, we did not want to curtail participant creativity. While submissions using external linguistic data did not qualify to win any of the tracks, they still qualified to be presented in the competition and to be published in the proceedings.

Community building. Given that the BabyLM Challenge aims to encourage research in efficient and cognitively plausible model pretraining, one of our goals was to encourage the formation of a research community with shared interests. Towards that end, we hosted a public messaging forum on Slack and enabled participants to interact with each other and with the task organizers. At the time of paper writing, this forum had over 250 members, including many interested researchers who did not ultimately submit to the challenge. An interactive forum was useful for both establishing a community and building interest; it allowed the community to clarify the track rules, debug the evaluation pipeline, and receive announcements from the organizers.

Timeline. Below, we replicate the timeline from the website.

- December 2022: The BabyLM Challenge is announced at CoNLL 2022, as well as on Twitter and in several mailing lists.
- January 2023: The pretraining datasets for the *Strict* and *Strict-Small* tracks were released.
- March 2023: The initial evaluation pipeline was made public.
- 1 June 2023: Hidden (surprise) evaluations were released and the Dynabench submission portal was opened.
- 22 June 2023: Deadline for model submissions (extended from 15 June 2023).
- 1 August 2023: Deadline for paper submissions.
- 6-7 December 2023: Presentation of the shared task at CoNLL.

4 Pretraining Corpus

We compiled and distributed a pretraining corpus inspired by the input received by children.[2] Submissions to the *Strict* track are required to train exclusively on this corpus. Submissions to the *Strict-Small* track are required to use only a scaled-down version of the dataset, approximately 10% the size of the *Strict*-track corpus. Two key properties of the dataset—its size and its domain—are controlled in order to make the data more developmentally plausible than typical LM pretraining data.

Size: 100M words or less. The pretraining corpus for the *Strict* track consists of under 100M words, and the corpus for the *Strict-Small* track is under 10M 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, which we round up to 100M words. The 10M word

[1] In our initial announcement, external software trained on linguistic input or expert annotations not included in our corpus—including taggers, parsers, tokenizers, or models were *not* allowed. However, numerous questions from participants prompted an announcement in April 2023 that we were modifying the rules of the *Loose* track to allow such methods. We made this decision because we determined that the interests of the community were better served by emphasizing creativity and discovery in the *Loose* track. Text generated by a language model that was trained only on a BabyLM corpus was not counted towards the 100M word budget, nor was data bootstrapped by such models.

[2] 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

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, 2020)	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/

Strict-Small dataset corresponds to the amount of input in the first two to five years of development. By contrast, contemporary widely used LMs such as Llama 2 (Touvron et al., 2023) are trained on trillions of words (Figure 1). Even BERT (Devlin et al., 2019), which is comparatively small by today's standards, was trained on over 3B words, well over the amount of input to a human in an entire lifetime. This discrepancy in input volume between LMs and humans is an oft-cited criticism of using these artifacts out-of-the-box as cognitive models (Warstadt and Bowman, 2022; Frank, 2023, a.o.).

Domain: Mostly transcribed speech. We source the majority ($\approx$ 56%) of the pretraining corpus from transcribed or scripted speech. We made this choice because the majority of the input to a hearing child comes from speech (though this proportion decreases with age as consumption of written media increases). This contrasts with standard LM training corpora, which consist mostly of text that was intended to be read and potentially edited. This is particularly significant for studying grammar learning, as some grammatical constructions (such as nominalizations and passives) are far more frequent in writing, while others (such as first- and second-person pronouns) are more frequent in speech (Biber, 1991).

Domain: Child-directed language. About 40% of the data in the pretraining corpus comes from sources either intended for children or appropriate for children, including child-directed speech, children's books, educational videos, and simplified English. Child-directed speech has been used as the sole or primary data source in some previous work aiming to model child language acquisition with LMs (Reali and Christiansen, 2005; Perfors et al., 2011; Pannitto and Herbelot, 2020; Huebner et al., 2021; Yedetore et al., 2023). We chose to include data from other domains (both child-directed and not) for several reasons. First, fewer than 10M words of transcribed child-directed speech are available, far below our 100M word budget. Second, child-directed speech makes up only part of the input to children. This amount can vary by a factor of 10 or more across cultures and socio-economic groups (Cristia et al., 2019). The estimate on which we base the 100M word budget (Gilkerson et al., 2017) counts *all* speech in the child's environment including overheard speech.

4.1 Contents

The contents of the BabyLM pretraining dataset are summarized in Table 1. Descriptions of each data source are provided in Appendix A.

4.2 Preprocessing

We release *Strict* and *Strict-Small* train, development, and test splits of each of the ten data sources, split approximately 83.3%/8.3%/8.3%. The 10M word *Strict-Small* training set is sampled randomly from the *Strict* training set. After any preprocessing, we downsample and split each source by randomly sampling chunks of 2000 lines or longer. The code and instructions for downloading and preprocessing the raw data are publicly available.[3]

We perform minimal preprocessing in terms of filtering and reformatting text. Notably, we gener-

[3]https://github.com/babylm/babylm_data_preprocessing.

ally preserve newlines in the original texts, meaning newlines do not consistently delimit documents, paragraphs, or sentences, as in some pretraining datasets. We use WikiExtractor (Attardi, 2015) to extract text from the `xml` Simple English Wikipedia dump dated 2022-12-01. We perform additional preprocessing on Simple English Wikipedia to remove `<doc>` tags. We select the spoken subset of the BNC by selecting only lines from the `xml` containing the `<stext>` tag and extracting only the text from the `xml`. We use code by Gerlach and Font-Clos (2020) to download and preprocess data from Project Gutenberg, which we additionally filter to contain only English texts by authors born after 1850. The OpenSubtitles and Wikipedia portions of the pretraining corpus were shared with us in preprocessed form, having had duplicate documents removed from OpenSubtitles and preprocessing steps performed to Wikipedia similar to our Simple English Wikipedia procedure.[4] We use regular expressions to remove speaker and dialog act annotations from the Switchboard Dialog Act Corpus. We perform no preprocessing on the remaining datasets.

5 Evaluation

To evaluate submissions, participants were asked to upload their model predictions to Dynabench, which is an online platform for dynamic data collection and model benchmarking.[5] Multiple submissions to the Dynabench platform were allowed, but at most one candidate was allowed to be chosen as a competitor from each team.

5.1 Evaluation Tasks

The goal of the evaluation pipeline is to assess the extent to which submitted models have learned the latent syntactic and semantic structure of their pretraining language. To evaluate the grammatical abilities of LMs, we use BLiMP (Warstadt et al., 2020a). BLiMP consists of tasks that evaluate the ability of language models to behave in a manner consistent with the structure of English. Each example consists of a minimal pair of sentences, where one sentence is acceptable and the other is unacceptable (differing as minimally as possible from the acceptable sentence otherwise); a model is correct on a given example if it assigns higher probability to the correct sentence in the minimal pair. We also release a supplement to the BLiMP tasks, which tests for phenomena not captured by BLiMP (see §5.1.1).

To assess the abilities of LMs on more typical downstream NLP tasks, we evaluate on a mixture of tasks from a subsample of (Super)GLUE, which consists of text classification tasks. We include a variety of task types, including paraphrase detection (MRPC, QQP), sentiment classification (SST-2), natural language inference (MNLI, QNLI, RTE), question answering (BoolQ, MultiRC), acceptability judgments (CoLA), and commonsense reasoning (WSC).

5.1.1 Hidden Tasks

Two weeks before the results deadline, we released three hidden evaluation tasks: the Mixed Signals Generalization Set (MSGS), a supplement to BLiMP, and an age-of-acquisition (AoA) prediction task. MSGS and the BLiMP supplement were mandatory; AoA prediction was provided as an additional analysis point for participants in writing their papers. The motivation for using these hidden tasks was to prevent our evaluations from rewarding submissions that overfit to the BLiMP and (Super)GLUE tasks.

The BLiMP supplement includes five test suites consisting of BLiMP-style minimal pairs that cover areas of linguistic knowledge not tested by BLiMP—namely, dialogue and questions. The test suites are semi-automatically generated using manually filled templates. As with BLiMP, models are evaluated on the supplement in a zero-shot manner, by comparing the probabilities of the sequences in a minimal pair, under the assumption that the acceptable sequence will be more probable than its unacceptable counterpart.

HYPERNYMS. We evaluate LMs' knowledge of lexical entailment, i.e., hypernym–hyponym relationships. This task bears similarity to natural language inference (Dagan et al., 2006; Bowman et al., 2015; Williams et al., 2018), but we instead measure whether models assign a higher likelihood to valid statements of entailment compared to minimally differing invalid statements. The evaluation data is designed around manually written triples consisting of ⟨`hypernym`, `base`, `hyponym`⟩—for example, ⟨*plant*, *herb*, *basil*⟩. We also specify an `other` noun (for example, *flower*) which shares the `hypernym` but not the `hyponym` with the `base` noun. From these nouns, plus a set of manually written

[4] We thank Haau-Sing Li for allowing us to use this preprocessed data.

[5] https://dynabench.org/

contexts, we generate six types of minimal pairs, shown in Table 5 in Appendix C. Additionally, we randomly vary the text used to convey entailment, e.g., *If p then q*, *If p that means q*, *p therefore q*, etc.

SUBJECT–AUXILIARY INVERSION. The subject–auxiliary inversion rule applies in question formation in English (e.g., relating *Logan will go.* to *Will Logan go?*). This task has been used to evaluate language models' syntactic abilities and preferences (e.g., McCoy et al., 2020; Mueller et al., 2022; Yedetore et al., 2023; Mueller and Linzen, 2023). Our test data was created by Warstadt (2022, Ch. 6), where it is described in more detail.

TURN-TAKING. Comprehending dialogue requires tracking the grammatical properties of utterances from multiple speakers. Pronouns such as *I*, *you*, and *she* are indexicals, meaning their interpretation depends on the speaker's context and identity. This test suite evaluates whether LMs can predict which pronoun is appropriate to use when there is a change in speaker. For example, if person A asks person B a question of the form *Can I ...*, person B's response should begin with *You*, not *I*. Our tests include (i) cases where the pronoun is expected to change, and (ii) cases where it is not. We also vary the context length (and therefore the distance between the context pronoun and the target), and whether the context contains a distractor pronoun in an embedded position. Finally, for each example, we randomly select one from a set of formats for indicating the speaker, e.g., *A: ..., B: ...*, or *"...," he asked. "...," she said.*, etc. Examples of each format can be found in Table 6 in Appendix C.

QUESTION–ANSWER CONGRUENCE. The syntax of a question constrains the acceptable responses. For example, a congruent answer to a *who*-question must be an animate noun (or contain one in a suitable context). This test suite evaluates whether LMs assign a higher likelihood to congruent answers compared to incongruent ones, and therefore learn the cross-sentential dependency between a *wh*-word and an answer. In addition to a set of EASY test cases, we construct a set of adversarial TRICKY test cases where there is a highly salient distractor answer that is not congruent with the *wh*-word. We randomly vary whether the answer appears as a fragment or in a complete sentence as well as the format for indicating the speaker. See Table 7 in Appendix C for examples.

Mixed Signals Generalization Set. The Mixed Signals Generalization Set (MSGS; Warstadt et al., 2020b) is a text classification task that evaluates the inductive biases of language models. For a MSGS subtask, models are finetuned on an ambiguous training set where the labels are consistent with both a syntactic generalization and a surface generalization, and then evaluated on examples that disambiguate which generalization the model converged on (if any).[6]

Ideally, models would be more sensitive to linguistic features than surface features, as a systematic preference for abstract linguistic properties allows models to generalize more robustly to unseen structures. The metric for MSGS is the Matthews correlation coefficient between the model's predictions and the labels according to the linguistic generalization on the test set. A coefficient of 1 corresponds to a systematic linguistic generalization, and -1 to a systematic surface generalization. Indeed, Warstadt et al. (2020c) find that linguistic bias increases with the volume of pretraining data, and that models with RoBERTa-like architectures require more than a billion words of pretraining data to achieve an overall linguistic bias (i.e., a score greater than 0).

Age-of-acquisition Prediction. Optionally, participants could evaluate on the age of acquisition (AoA) prediction task of Portelance et al. (2023). When humans are learning language, they tend to acquire certain words at specific ages; the age of acquisition of a word refers to the age at which humans acquire that word. The AoA prediction task compares LMs' word surprisals with children's AoA of the same words. A language model's average surprisals are converted into AoA predictions, and these are then compared to the actual average AoA (in months) of those words. Models achieving lower mean absolute deviation between the actual

[6]For example, one of the subtasks tests which of the following two generalizations the model's inductive bias favors: whether the word "the" is present (the surface generalization), or whether the sentence contains an adjective (the syntactic generalization). Thus, training examples will include only ambiguous labeled pairs where these two properties are both perfectly correlated with each other and with the binary labels, such as `(The big dog barked, 1)` and `(A dog barked, 0)`. At test time, the model must classify held-out sentences where the features are anti-correlated, such as `A big dog barked` and `The dog barked`. If the model predicts labels `1` and `0` respectively for these and other analogous examples, we infer that it classifies examples based on the linguistic feature, while if it predicts `0` and `1` respectively, it adopted the surface generalization.

age and predicted age are said to perform better on the task.[7] While we did not require participants to submit these scores as part of their predictions, we provided code to make evaluation on this task simple, such that they could include this score as an additional analysis point in their paper submissions. 7 teams (22.6%) evaluated on the AoA prediction task; see Appendix E for results and discussion.

5.2 Evaluation Pipeline

The organizers provided code to unify the evaluation setup across submissions. This was released as a public repository on GitHub.[8] The evaluation pipeline supports models implemented in HuggingFace, though we did not restrict the model submissions to HuggingFace-based models.[9] For model and result submissions, users were required to (i) upload a link to their model (on any file-hosting service), and (ii) provide model predictions for each example of each task (via Dynabench); we provided a template specifying the format of the predictions file.

Data preprocessing. NLP tasks in our evaluation pipeline often contained vocabulary that is not contained in the BabyLM pretraining corpora. To address this mismatch, we filtered each task according to its lexical content: if an example contained any words that appear less than twice in the *Strict-Small* training corpus, we filtered the example out. Otherwise, each dataset is presented in its original format. See Table 4 in Appendix B for details on the size of the filtered datasets.

5.2.1 Evaluation Paradigms

Zero-shot evaluation. For zero-shot tasks—BLiMP and the BLiMP supplement—we modify the BigScience fork of the `lm-eval-harness` repository, originally by EleutherAI (Gao et al., 2021). This provides functionality for scoring autoregressive decoder-only LMs and encoder-decoder LMs. For encoder-only LMs, we modify the repository to support masked language model scoring as described in Salazar et al. (2020).[10]

Finetuning. We first attempted zero-shot learning and few-shot in-context learning for (Super)GLUE and MSGS tasks. However, this often resulted in random-chance accuracies from each of our baselines; we, therefore employ finetuning.[11] For tasks requiring finetuning—(Super)GLUE (Wang et al., 2018, 2019) and MSGS (Warstadt et al., 2020b)—we base our scripts on HuggingFace's example finetuning scripts for text classification.[12] We modified the script to support encoder-decoder models, and to work for a wider variety of tasks. We provide a default set of hyperparameters that we found to work well across our baseline models, though participants were allowed to freely modify hyperparameters.

5.3 Dynabench Leaderboard

Dynabench is an open-source platform for dynamic dataset creation, model evaluation, and leaderboard hosting (Kiela et al., 2021). In addition to open-sourcing datasets—including adversarial and human-in-the-loop datasets (Nie et al., 2020; Bartolo et al., 2021; Potts et al., 2021; Sheng et al., 2021; Vidgen et al., 2021; Kirk et al., 2022)—Dynabench has offered leaderboard support for several community challenges in the past (Wenzek et al., 2021; Bartolo et al., 2022; Mazumder et al., 2022). Given that we desire a dynamic leaderboard that allows for submissions even after the end of the challenge, this platform was well-suited to the BabyLM Challenge. All model submissions to the challenge were submitted via the Dynabench platform, to the respective leaderboards for the *Strict*,[13] *Strict-Small*,[14] and *Loose*[15] tracks.

Each leaderboard presents aggregate scores across all tasks, which can be interactively bro-

[7] It is not clear whether optimizing LM performance on this task necessarily leads to better language models. It is possible instead that LMs could have a different pattern of surprisals than humans while learning particular linguistic concepts more or less efficiently than humans. Thus, this task should be used more as a measure of how well LMs align with humans—and thus, as a measure of their usefulness as cognitive models of language acquisition and processing—rather than as a measure of quality or performance.

[8] https://github.com/babylm/evaluation-pipeline

[9] Upon release of the evaluation pipeline, we announced that we would provide support as needed to teams training LMs not based in HuggingFace.

[10] We use the implementation of Misra (2022) in the `minicons` library.

[11] finetuning technically adds to the training set size. We consider this acceptable, as finetuning on a single GLUE or MSGS task does not meaningfully add to the domain-general linguistic abilities of language models. The LM is finetuned separately for each task, so we still see this as an evaluation of the LM's abilities in itself (albeit more confounded than the zero-shot evaluations).

[12] https://github.com/huggingface/transformers/blob/211f93aab95d1c683494e61c3cf8ff10e1f5d6b7/examples/pytorch/text-classification/run_glue.py

[13] https://dynabench.org/tasks/baby_strict

[14] https://dynabench.org/tasks/baby_strict_small

[15] https://dynabench.org/tasks/baby_loose

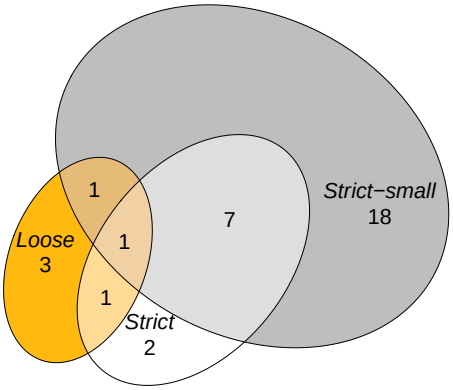

Figure 2: Number of participants who submitted to each track, with multiple submissions counted once.

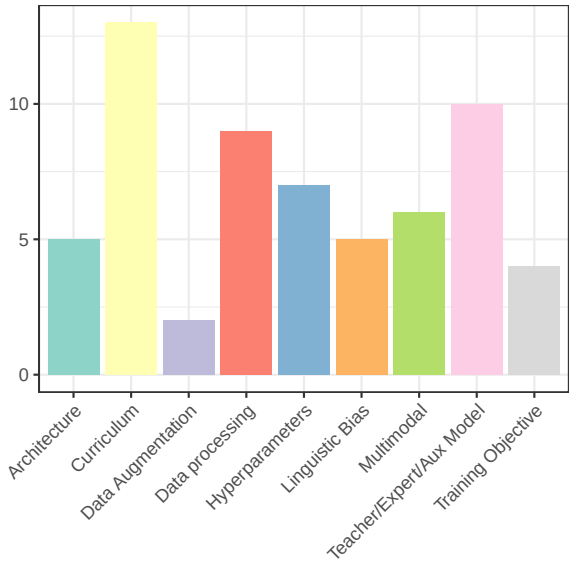

Figure 3: Total number of submitted models that used each of the nine approaches in our typology. We count at most one submitted model per participant per track.

ken down into more fine-grained scores per task and per subtask. To compute the aggregate score, we weigh BLiMP and the BLiMP-supplement together at 50% (all subtasks weighted equally), (Super)GLUE at 30%, and MSGS at 20%. This weighting scheme was arrived at heuristically, though we did observe that the winners for each track were stable across a wide range of reasonable weightings. Dynabench allows users to specify a custom task weighting to compute an alternative aggregate score. The leaderboard for the BabyLM challenge will continue to accept submissions indefinitely.

5.4 Baselines and Skylines

Baselines. To provide simple baselines for our evaluation tasks, we train multiple models on the data released for *Strict-Small* and *Strict* tracks and evaluate them on the evaluation tasks. Three baseline models are provided: OPT-125M, RoBERTa-base, and T5-base. These models use the same objective function and network architecture corresponding to their original papers (OPT; Zhang et al., 2022, RoBERTa; Liu et al., 2019, T5; Raffel et al., 2020). The network architecture of these models covers both encoder-decoder (T5-base and RoBERTa-base) and decoder-only (OPT-125M) architectures. Their objective functions include next-token prediction (OPT-125M), masked-token prediction (RoBERTa-base), and sequence-to-sequence (T5-base) matching losses. The baseline models are trained using a fixed context length of 128, a constant learning rate of 1e-4, a linear learning-rate warmup from 0 in the first 5000 steps, a batch size of 128, and AdamW (Loshchilov and Hutter, 2019) as the optimizer. They are trained for 20 epochs on the data, where each epoch randomly and independently shuffles the whole dataset. Although most of these hyperparameters are loosely inspired by Huebner et al., we expect that the specific choices on them can be further improved and leave these potential improvements as possible topics for submissions. We find that our baseline models achieve reasonable performance on the evaluation tasks, with clear improvement from more data from *Strict-Small* to *Strict* track and notable gap towards their counterparts pretrained on much larger datasets.

Skylines. To get an approximation of how well larger models could, in principle, perform in our task and setting, we ran Llama 2 70B (Touvron et al., 2023) and the fully trained RoBERTa-base model through our evaluation pipeline. This is meant to provide a comparison point to the state of the art in 2023, as the Llama 2 model is pretrained on much more data (2T tokens) than the challenge allows, and it has far more parameters than we expect to find in submissions. We evaluate Llama 2 on (Super)GLUE using in-context learning, but it is fully finetuned on MSGS. BabyLM submissions that approach these scores can be considered to have greater sample efficiency than the skyline models, and may therefore provide stronger starting points for future research in sample-efficient NLP.

6 Submissions Summary

We received 31 papers and 162 models in total. Table 3 shows the submission counts for each track.

	Model	BLiMP	GLUE	MSGS	BLiMP-Supp.	*Aggregated*
	Llama 2	0.84	0.84	0.26	0.75	0.71
	RoBERTa-Base	0.87	0.79	0.24	0.76	0.70
Strict	ELC-BERT (Charpentier and Samuel, 2023)	0.85	0.78	**0.47**	**0.77**	**0.74**
Strict	BootBERT (Samuel, 2023)	**0.86**	**0.79**	0.28	0.72	0.70
Strict	McGill-BERT (Cheng et al., 2023)	0.84	0.72	0.25	0.71	0.67
Strict	*Best Baseline (OPT-125M)*	0.75	0.70	0.13	0.68	0.60
Strict-Small	ELC-BERT (Charpentier and Samuel, 2023)	**0.80**	**0.74**	**0.29**	0.67	**0.66**
Strict-Small	MLSM (Berend, 2023b)	0.79	0.71	0.17	0.57	0.61
Strict-Small	McGill-BERT (Cheng et al., 2023)	0.75	0.70	0.13	**0.68**	0.60
Strict-Small	*Best Baseline (OPT-125M)*	0.63	0.62	0.10	0.53	0.50
Loose	Contextualizer (Xiao et al., 2023)	**0.86**	**0.73**	**0.58**	0.63	**0.73**
Loose	McGill-BERT (Cheng et al., 2023)	0.80	0.68	-0.02	0.57	0.57
Loose	BabyStories (Zhao et al., 2023)	0.78	0.61	0.03	**0.65**	0.56

Table 2: Top 3 systems for each track, as well as the baseline model with the highest aggregate score. We also show "skyline" models: RoBERTa-base and Llama 2 trained on their full pre-training corpora. Each task score is simply the mean score across each of its subtasks. The aggregate score is a weighted average of each task. We **bold** the highest-scoring system for each task within each track.

	# Models	# Participants
Loose	20	8
Strict-Small	118	29
Strict	24	11
total	*162*	*31*

Table 3: Total number of models and participants per track. Participants who submitted to multiple tracks are counted once in the total.

Some participants submitted to multiple tracks; we show data for unique participants in Figure 2.

We found that many submissions focused their efforts on similar techniques. To better quantify this, we devised a typology of the nine most common approaches and assigned each submitted model one or more labels. Figure 3 shows the number of submissions employing each approach. §7.3 provides more detailed descriptions of each approach, as well as results indicating which ones were most effective.

All participants are affiliated with universities or independent research institutions. Participants' home institutions are located in 16 different countries. The number of participants by country is as follows (multinational participants are counted more than once): US (9), Germany (5), Netherlands (3), UK (4), Canada (2), Norway (2), Austria (1), Denmark (1), France (1), Hungary (1), Israel (1), Japan (1), Norway (1), Switzerland (1), Turkey (1).

The official leaderboard is available on Dynabench.[16] With the consent of participants, we release links to submitted models, their complete predictions for the evaluation tasks, their scores for each task and subtask, and metadata about each submission at the BabyLM's GitHub at https://github.com/babylm/submissions2023. We provide a summary of each submission in Appendix F.

7 Results & Analysis

7.1 Overall Results & Track Winners

The results from all submissions are shown in Figure 4, with the scores of the top-performing models in each track detailed in Table 2. In the figure, dashed green lines show the performance of the Llama 2 skyline. Solid green lines show human performance on GLUE reported in Nangia and Bowman (2019), and human performance on BLiMP as reported by Warstadt et al. (2020a).

Before discussing the winning systems in each track, we note a few high-level takeaways from these results. The strongest results were achieved by models in the *Strict* track. Given the *Strict* track's larger training corpus relative to the *Strict-Small* corpus, it is not surprising that these models could outperform those in the *Strict-Small* track. However, there are two interesting trends: First, *Strict* models did not outperform those in *Strict-Small* by a large amount, even though the size of training data was an order-of-magnitude larger. For example, there are only

[16] https://dynabench.org/babylm

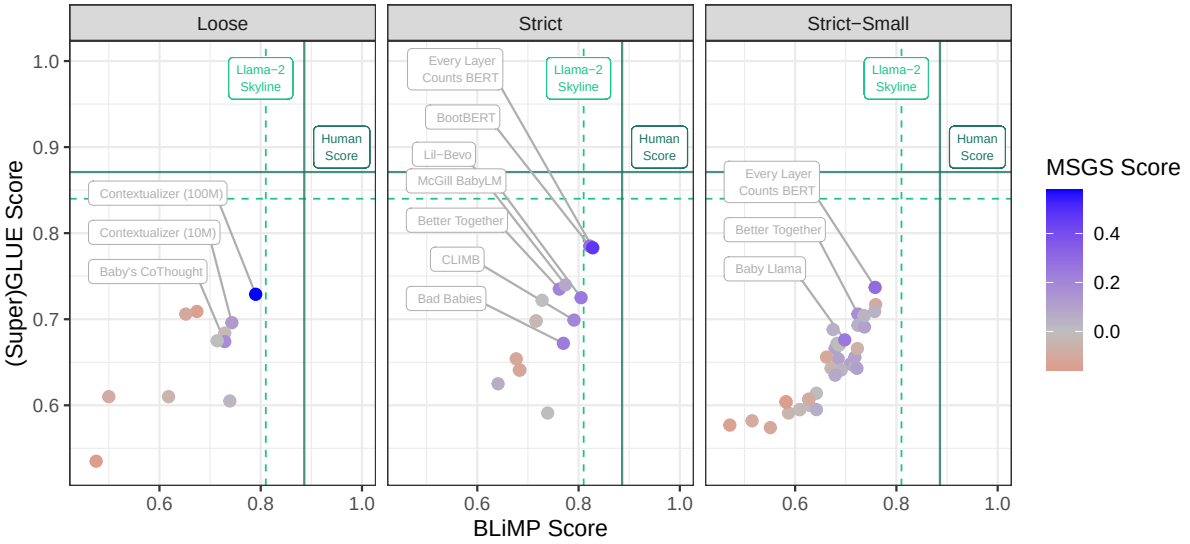

Figure 4: **Summary of BabyLM Submission Results:** Each point represents an official model submission. Scores are broken down into performance on BLiMP (x-axis), GLUE (y-axis) and MSGS (color). Submissions that achieve an aggregate score above 0.6 are labeled in gray. Green dashed lines show Llama 2 skyline performance, and green solid lines show the human performance ceiling.

two models in the *Strict* track that achieve higher GLUE scores than the best-performing *Strict-Small* model. Second, models in the *Loose* track tended to perform worse in the aggregate than those in the *Strict-Small* track, even though they potentially had access to additional (non-linguistic) data. One conclusion we can draw from this is that learning from multiple modalities of data presents a challenge in its own right, and that current model architectures are not optimized to efficiently utilize multiple types of inputs during training.

The other important high-level takeaway is that many BabyLM models are very close to the Llama 2 skyline, and to achieving human-level performance on BLiMP and GLUE (i.e., they are near the green lines in Figure 4). Strong performance could be expected in the case of (Super)GLUE, where models were finetuned with additional data, but we note that even for BLiMP, the top-performing model is only about 3% shy of human performance. Note that prior to the start of the challenge, we explored the possibility of measuring zero-shot performance on (Super)GLUE test sets, and found zero-shot performance to be at or below chance for our baselines. This fact, as well as the consideration that GLUE has been traditionally evaluated using finetuning, leads us to select finetuning evaluations for the (Super)GLUE benchmark(s).

Given that successful training on developmentally plausible corpora could have ramifications for cognitive and linguistic theories of learnability (Wilcox et al., 2023; Warstadt and Bowman, 2022), these results point to two important takeaways: (1) Human-level results have not been achieved *yet*. However, (2) given the strong performance of the top-scoring models, human-level results appear likely to be achieved very soon, possibly within the next few years. Of course, one possible concern is the following: current models may not be close to human-level performance; rather, current performance metrics, like BLiMP, might not accurately measure human-level linguistic competence. We are sympathetic to such concerns, but we also note that BLiMP, and other related syntactic benchmarks such as those presented in Marvin and Linzen (2018) and Gauthier et al. (2020), were specifically designed to mimic the types of tests invented by linguists and cognitive scientists to reveal syntactic competence—i.e., they are all based on minimal pair sentences. Thus, while it is imperative to continue building more comprehensive and larger datasets, we believe it is fair to say that the close-to-human scores observed in the BabyLM challenge on BLiMP reflect genuine grammatical generalizations learned by the models.

7.2 Winning Submissions

Below, we discuss the winning submissions from each track in greater detail. We also mention the winners of our "Most Interesting Paper" awards

and provide a brief justification for each.

***Strict* track.** The winner of the *Strict* track is ELC-BERT submitted by Charpentier and Samuel (2023). This model, as well as the runner-up submission Boot-BERT (Samuel, 2023), used as their starting point the LTG-BERT architecture from Samuel et al. (2023). Although these submissions make additional incremental improvements to the LTG-BERT training regime, their own baselines suggest that the backbone architecture plays a large role in the submissions' successes. LTG-BERT's main contribution is a synthesis of several optimizations to the Transformer architecture, namely: (1) additional layer normalization, following (Shleifer et al., 2021); (2) GEGLU feed-forward modules (Shazeer, 2020); (3) disentangled attention following DeBERTa (He et al., 2021); and (4) scaled weight initialization following (Nguyen and Salazar, 2019). ELC-BERT modifies this backbone such that the input to each layer is a weighted sum of the outputs of all previous layers. Another notable property of LTG-BERT is that all models with this architecture so far have been trained for a large number of epochs. Charpentier and Samuel (2023) train models for over 450 epochs for their *Strict* submission, and over 2000 epochs for their *Strict-Small* submission. LTG-BERT models performed exceptionally well on our set of evaluations, outperforming not only every other submission to the shared task but also the Llama 2 and RoBERTa-Base skylines on overall score and on all test suites except for (Super)GLUE (Table 2). The second runner-up for this track was McGill-BERT (Cheng et al., 2023).

***Strict-Small* track.** The winner of the *Strict-Small* track is, again, ELC-BERT (Charpentier and Samuel, 2023). This double-win demonstrates that the model's architectural choices work well with multiple scales of pretraining data. The runners-up were MLSM (Berend, 2023b) and McGill-BERT (Cheng et al., 2023).

***Loose* track.** The winner of the *Loose* track is the Contextualizer model of Xiao et al. (2023), which used a data processing scheme in which extra training samples are created by combining chunks of texts from different contexts. Repeating this process 40 times for each chunk gives a dataset that has as many training samples as 4B word dataset, but based on a dataset of only 100M words. This augmentation technique outperforms training for 40 epochs using the same training samples. Runners-up for this track were McGill-BERT (Cheng et al., 2023) and the BabyStories model of Zhao et al. (2023).

Most interesting paper awards. These awards are given to papers that go beyond achieving high scores on a leaderboard, and instead demonstrate contributions to the shared task based on interesting analyses, useful negative results, creative modeling choices, or a combination thereof. We awarded two most interesting paper awards in two different categories.

Outstanding evaluation. The most interesting paper award for outstanding evaluation was given to "Large GPT-like Models are Bad Babies: A Closer Look at the Relationship between Linguistic Competence and Psycholinguistic Measures" (Steuer et al., 2023). This work goes beyond the BabyLM evaluation tasks: the authors use measures of human cognitive processing effort and linguistic competence and additionally correlate these with BabyLM task performance. Their work assesses BabyLM submissions as models of human language processing, thus contributing to our understanding of how to better train cognitive models.

Compelling negative results. The most interesting paper award for compelling negative results was given to "CLIMB—Curriculum Learning for Infant-inspired Model Building" (Martinez et al., 2023). This work proposes a typology of common curriculum learning approaches and performs a thorough and principled evaluation exploring this design space. Although they find that none of the tested approaches leads to widespread improvements across the evaluation tasks, the exhaustiveness of this search and the careful controls and baselines in the study make this negative result a valuable contribution.

7.3 Common Methods

One of the main objectives of the BabyLM Challenge is to compare and contrast methodological choices for sample-efficient pretraining. To do so, we hand-coded each submission based on the method(s) it employs. Figure 3 shows the number of submissions using each approach, and we visualize the performance of different methods in Figure 5. We also present a similar figure separated by the underlying architecture (Figure 6). Each of these approaches is discussed

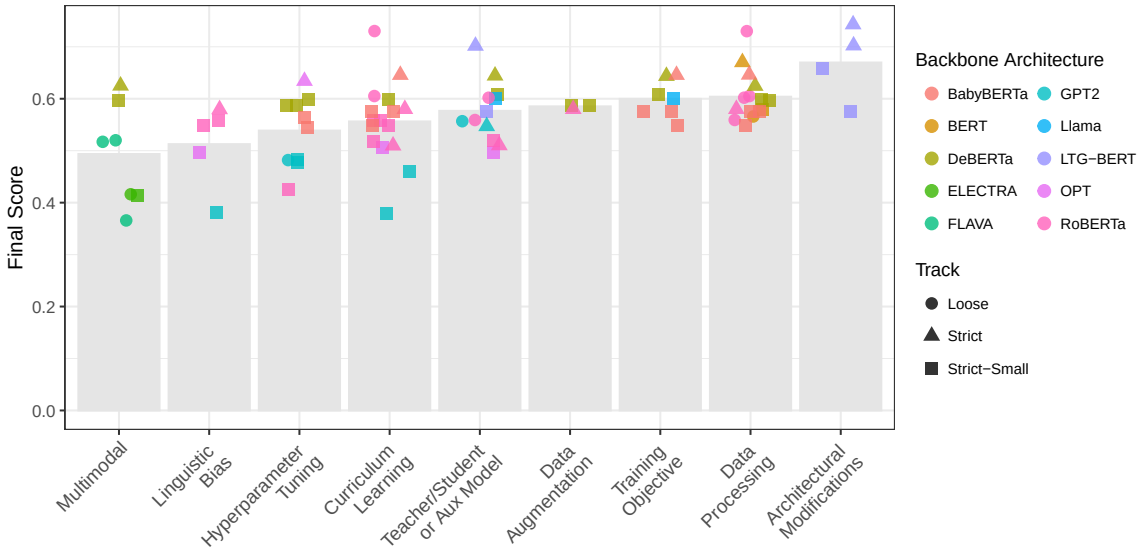

Figure 5: **Effect of Training Strategy and Backbone Architecture:** Each point represents a submission. Some submissions may appear more than once if they use multiple strategies. Shapes show the challenge track to which the model was submitted. Colors show the backbone architecture on which the model is based. Gray bars show within-category aggregates.

in further detail below. We highlight two high-level takeaways to start: First, curriculum learning, which was the most popular approach, did not tend to produce high scores (although one curriculum learning model did perform well). Second, the highest-performing models were ones that made architectural modifications—namely, those based on the LTG-BERT architecture.

Curriculum learning. This approach entails sorting training steps with respect to some complexity metric(s). This was the most popular approach, with 13 teams (41.9%) attempting some variant of curriculum learning. The majority of these attempts did not produce consistent improvements across the BabyLM evaluation tasks. However, they did explore a large space of possible curricula, for example: ranking sentences by surprisal (Chobey et al., 2023; Hong et al., 2023), lexical frequency (Borazjanizadeh, 2023; Martinez et al., 2023), length (DeBenedetto, 2023; Edman and Bylinina, 2023), and syntactic complexity (Mi, 2023; Oba et al., 2023; Bunzeck and Zarrieß, 2023); sorting entire datasets by difficulty (Opper et al., 2023; Martinez et al., 2023; Xiao et al., 2023); gradually increasing vocabulary size (Thoma et al., 2023; Edman and Bylinina, 2023); and gradually increasing the difficulty of the training objective (Martinez et al., 2023).

Teacher–student or auxiliary model. Many papers trained their submitted models with the aid of additional models. According to our rules, this was permissible as long as any auxiliary models were trained on the BabyLM corpus. Knowledge distillation using auxiliary models was often a successful approach: Samuel (2023) considered an exponential moving average teacher model (Tarvainen and Valpola, 2017), while Berend (2023b) modeled a latent semantic feature distribution from a teacher model. Timiryasov and Tastet (2023) performed distillation on an ensemble of features. Others used auxiliary models to select appropriate training examples for a curriculum (Chobey et al., 2023; Hong et al., 2023), or trained a reward model for use in reinforcement learning (Zhao et al., 2023).

Data preprocessing. Many submissions modified the format of the pretraining corpus. When controlled comparisons were performed, these preprocessing steps often led to improvements. In §7.2 we discuss the successful Contextualizer method for constructing new training samples. Other successful approaches used short sequences or individual sentences as training samples, rather than long portions of documents (Govindarajan et al., 2023; Cheng et al., 2023; Edman and Bylinina, 2023). Among the more unique approaches in this space was Baby's CoThought (Zhang et al., 2023), which used an LLM to reformat unrelated sentences from the corpus into coherent paragraphs.

Hyperparameter tuning and model scaling. This was a relatively common approach. Many

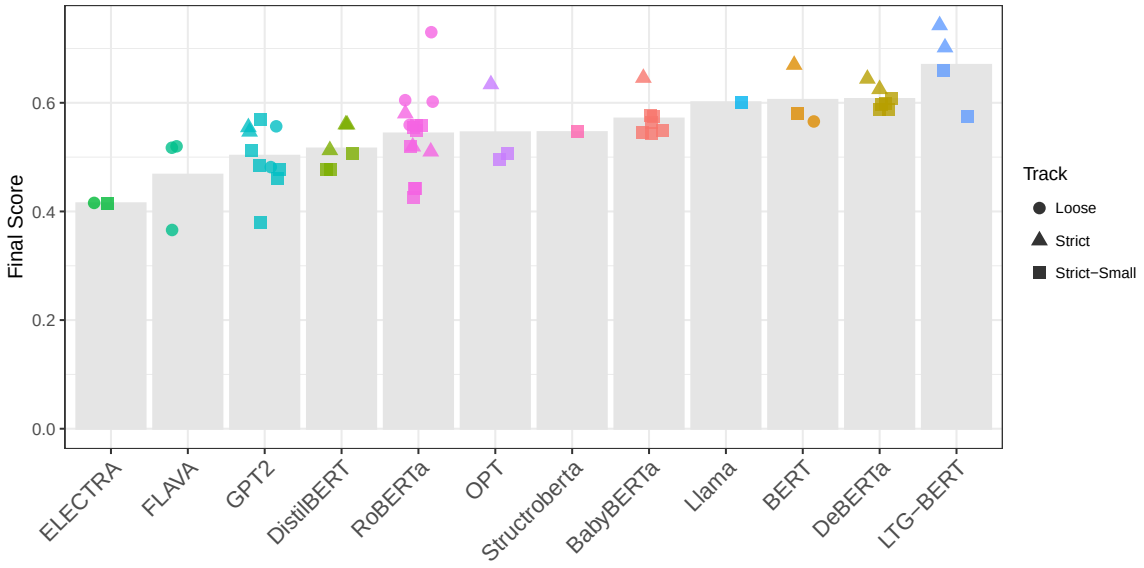

Figure 6: **Effect of Backbone Architecture:** Each point represents a submission. Shape indicates the challenge track. Gray bars show within-category aggregates.

submissions performed extensive hyperparameter searches, producing hard-won hyperparameters that work well on smaller datasets while preserving features of the dataset. While extensive hyperparameter searching can be expensive and challenging when scaling up to full-sized pretraining, in our limited data regime, consistently successful modifications include reducing context length (see "Data preprocessing", above), and training for more epochs or long epochs with data augmentation (Jumelet et al., 2023; Bhardwaj et al., 2023; Yang et al., 2023; Xiao et al., 2023; Samuel, 2023; Charpentier and Samuel, 2023).

However, results are mixed when modifying model size: some participants achieved better results when scaling model sizes up (Çağatan, 2023), while others were able to perform well when using very small models (Proskurina et al., 2023). More controlled studies using a variety of architectures and datasets are needed to determine whether scaling up or down is a better solution.

Multimodal learning. Multimodal learning was one of the directions where we expected the most interest and the most submissions; however, we received few submissions based on multimodal inputs, and the multimodal submissions did not reliably contribute to higher overall accuracy. One submission used music (Govindarajan et al., 2023), another used vision and language data (Amariucai and Warstadt, 2023), a third explored text-and-audio (Wolf et al., 2023), and a fourth incorporated text-and-image data and lexical sensorimotor data as part of the embedding process using multiplex networks (Stella et al., 2017; Ciaglia et al., 2023). Music training produced minor improvements on some subtasks, while the vision-and-language system marginally improved over the baselines in the *Strict-Small* track. The multiplex network did not produce performance gains, though it did allow the participants to reduce the number of parameters while preserving performance relative to the baselines. WhisBERT was reported to be undertrained, making its results difficult to interpret.

Architecture modifications. The winning submission made architectural modifications: Charpentier and Samuel (2023) made slight improvements to LTG-BERT (see §7.2 for more on this architecture) by taking a weighted sum over the outputs of all previous layers. Momen et al. (2023) used the relatively novel StructFormer architecture (Shen et al., 2021), which encourages tree-structured representations of inputs.

Training objectives. Some submissions trained language models using a mixture of both a language modeling objective and some other objective. Knowledge distillation from teacher models (see paragraphed titled "Teacher–student or auxiliary model" above) was the most common modification. Martinez et al. (2023) simplified the masked language modeling objective by coarse-graining the output classes, with little effect. Govindarajan et al. (2023) achieved improvements on specific BLiMP subtasks by modifying the masking procedure to preferentially mask specific words thought to be rel-

evant to a particular phenomenon tested by BLiMP.

Linguistic bias. Some submissions tried to impart human linguistic biases to models. Such approaches discussed above include curriculum learning based on linguistically motivated data sorting methods and architectures like StructFormer that encourage hierarchical analyses of inputs. Chen and Portelance (2023) also pretrained with token embeddings obtained via grammar induction, and Thoma et al. (2023) iteratively updated the vocabulary of the LM based on word simplicity measures (motivated by human age-of-acquisition analyses).

Data augmentation. Arguably, the effective Contextualizer approach (Xiao et al., 2023) is a form of data augmentation (see §7.2). Jumelet et al. (2023) used regular expressions to generate question-answer pairs given the BabyLM training data. Zhao et al. (2023) used an LLM to generate text merging disparate sentences into cohesive paragraphs.

8 Future BabyLM Challenges

The first iteration of the BabyLM Challenge yielded many successes, but also some organizational and scientific challenges. The lessons learned from our findings can improve future iterations of this challenge.

We were surprised that there were significantly more submissions to the *Strict-Small* track than the other two tracks combined, considering that the *Loose* track allows for a much wider variety of methods. However, this is understandable from the perspective of compute: training on *Strict-Small* is the least computationally expensive of each of the tracks, and it constrains the model search space enough that ideas are perhaps easier to define and execute. In future iterations of the BabyLM challenge, it could be interesting to provide more specific and constrained *Loose* tracks, which focus on particular research directions—for example, LLM-assisted low-resource pretraining, allowing expert annotations during pretraining, or joint text and audio modeling.

We can also draw insights from the data preprocessing and hyperparameter tuning submissions in particular, and standardize them into the dataset/evaluation pipeline. For example, we could preprocess the data in ways the present challenge has shown to be effective. This could include sorting the data according to the curriculum learning method that yielded performance gains, providing better-starting hyperparameters, and training a baseline with the best architecture.

Although data quantity was the main focus of this iteration, we may also consider rewarding compute efficiency in the future. Many of the most successful submissions consumed a lot of compute by training for many epochs. Indeed, the winning submission trained on about as many samples as BERT, despite having a training set only about 3% as large. While this finding is interesting, it does little to help achieve our goals in §2. Training for hundreds of epochs is not cognitively plausible, and it is does not make it easier and more accessible to test novel training approaches or train models on a university budget.

The evaluation pipeline was built on the existing `lm-evaluation-harness` repository,[17] but maintaining and updating it for this challenge was no small feat for a single organizer. In future iterations of the challenge, it would be beneficial to have a larger dedicated support team for the evaluations. A dedicated team could also allow us to handle a greater variety of submissions, including those not supported by HuggingFace.

9 Conclusions

The BabyLM Challenge encouraged participants to *think small*. We asked: can we improve language modeling on smaller and more cognitively plausible datasets? The submitted systems employed diverse methods, but the most consistent gains came from modified model architectures, new training objectives, principled preprocessing of the pretraining corpora, and hyperparameter searches. In one case, a curriculum learning method resulted in significant improvements. Future work can build on these findings to further improve language modeling for low-resource settings and for cognitive modeling research.

Acknowledgments

We would like to thank the participants of the BabyLM Challenge for their valuable contributions—not just their models and papers, but also their contributions to the evaluation pipeline and the reviewing process.

[17] Originally released at https://github.com/EleutherAI/lm-evaluation-harness. Note that we based our implementation on the BigScience fork at https://github.com/bigscience-workshop/lm-evaluation-harness.

We would also like to thank the Dynabench team at MLCommons for hosting our leaderboards and integrating our challenge's unique requirements into their implementation. Thanks especially to Max Bartolo, Douwe Kiela, and Hannah Rose Kirk for feedback on earlier iterations of the BabyLM evaluation setup.

Author Contributions

- **Original concept:** Alex Warstadt, Leshem Choshen
- **Primary organizers:** Alex Warstadt, Ethan Wilcox, Leshem Choshen, Aaron Mueller, Chengxu Zhuang
- **Pipeline implementation and maintenance:** Aaron Mueller
- **Baseline model training**: Chengxu Zhuang
- **Publicity and communications with participants:** Leshem Choshen, Ethan Wilcox
- **Training dataset compilation:** Alex Warstadt
- **BLiMP Supplement evaluation data creation:** Alex Warstadt
- **Dynabench integration:** Juan Ciro, Rafael Mosquera, Adina Williams
- **Llama 2 evaluation:** Bhargavi Paranjape
- **Guidance on concept and workshop organization:** Ryan Cotterell, Tal Linzen, Adina Williams
- **Reviewing submissions:** Alex Warstadt, Ethan Wilcox, Leshem Choshen, Aaron Mueller, Chengxu Zhuang, Adina Williams
- **Initial draft of findings paper:** Alex Warstadt, Ethan Wilcox, Leshem Choshen, Aaron Mueller, Chengxu Zhuang
- **Editing:** All authors

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*, pages 1856–1862, Reykjavik, Iceland. European Language Resources Association (ELRA).

Theodor Amariucai and Alexander Scott Warstadt. 2023. Acquiring linguistic knowledge from multimodal input. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic Structures in Natural Language*, pages 1–16.

Max Bartolo, Hannah Kirk, Pedro Rodriguez, Katerina Margatina, Tristan Thrush, Robin Jia, Pontus Stenetorp, Adina Williams, and Douwe Kiela, editors. 2022. *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*. Association for Computational Linguistics, Seattle, WA.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Eden Bensaid, Mauro Martino, Benjamin Hoover, Jacob Andreas, and Hendrik Strobelt. 2021. Fairytailor: A multimodal generative framework for storytelling. *CoRR*, abs/2108.04324.

Gábor Berend. 2023a. Masked latent semantic modeling: an efficient pre-training alternative to masked language modeling. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13949–13962, Toronto, Canada. Association for Computational Linguistics.

Gábor Berend. 2023b. Better together: Jointly using masked latent semantic modeling and masked language modeling for sample efficient pre-training. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Khushi Bhardwaj, Raj Sanjay Shah, and Sashank Varma. 2023. Pre-training LLMs using human-like development data corpus. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Douglas Biber. 1991. *Variation across Speech and Writing*. Cambridge University Press.

Nasim Borazjanizadeh. 2023. Optimizing GPT-2 pretraining on BabyLM corpus with difficulty-based sentence reordering. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Ellen Breitholtz, Shalom Lappin, Sharid Loaiciga, Nikolai Ilinykh, and Simon Dobnik, editors. 2023. *Proceedings of the 2023 CLASP Conference on Learning with Small Data*. Association for Computational Linguistics, Gothenburg, Sweden.

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.

Bastian Bunzeck and Sina Zarrieß. 2023. GPT-wee: Effective pre-training for downsized language models. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Margaret F. Carr, Shantanu P. Jadhav, and Loren M. Frank. 2011. Hippocampal replay in the awake state: a potential substrate for memory consolidation and retrieval. *Nature Neuroscience*, 14(2):147–153.

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*. Association for Computational Linguistics (ACL).

Xuanda Chen and Eva Portelance. 2023. Grammar induction pretraining for language modeling in low resource contexts. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Ziling Cheng, Rahul Aralikatte, Ian Porada, Cesare Spinoso-Di Piano, and Jackie C. K. Cheung. 2023. McGill BabyLM shared task submission: The effects of data formatting and structure biases. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Aryaman Chobey, Oliver Smith, Anzi Wang, and Grusha Prasad. 2023. Can training neural language models on a curriculum with developmentally plausible data improve alignment with human reading behavior? In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Leshem Choshen, Elad Venezian, Shachar Don-Yehiya, Noam Slonim, and Yoav Katz. 2022. Where to start? analyzing the potential value of intermediate models. *CoRR*, abs/2211.00107.

Floriana Ciaglia, Massimo Stella, and Casey Kennington. 2023. Investigating preferential acquisition and attachment in early word learning through cognitive, visual and latent multiplex lexical networks. *Physica A: Statistical Mechanics and its Applications*, 612:128468.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *Proceedings of the 8th International Conference on Learning Representations*. OpenReview.net.

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive.

Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. 2019. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, 90(3):759–773.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*. Springer.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Justin DeBenedetto. 2023. Byte-ranked curriculum learning for BabyLM strict-small shared task 2023. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Diskin, Alexey Bukhtiyarov, Max Ryabinin, Lucile Saulnier, Quentin Lhoest, Anton Sinitsin, Dmitry Popov, Dmitriy Pyrkin, Maxim Kashirin, Alexander Borzunov, Albert Villanova del Moral, Denis Mazur, Ilia Kobelev, Yacine Jernite, Thomas

Wolf, and Gennady Pekhimenko. 2021. Distributed deep learning in open collaborations. In *Advances in Neural Information Processing Systems*.

Shachar Don-Yehiya, Elad Venezian, Colin Raffel, Noam Slonim, and Leshem Choshen. 2023. ColD fusion: Collaborative descent for distributed multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 788–806, Toronto, Canada. Association for Computational Linguistics.

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

Lukas Edman and Lisa Bylinina. 2023. Too much information: Keeping training simple for BabyLMs. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211. Wiley Online Library.

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*. Association for Computational Linguistics (ACL).

Michael C. Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 22(1). Number: 126 tex.pubmedid: 33285901.

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.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1.

Venkata Subrahmanyan Govindarajan, Juan Diego Rodriguez, Kaj Bostrom, and Kyle Mahowald. 2023. Lil-bevo: Explorations of strategies for training language models in more humanlike ways. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations*.

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, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Xudong Hong, Sharid Loáiciga, and Asad B. Sayeed. 2023. A surprisal oracle for active curriculum language modeling. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

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.

Philip A. Huebner and Jon A. Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In Kara D. Federmeier and Lili Sahakyan, editors, *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of learning and motivation*, pages 279–331. Academic Press. ISSN: 0079-7421.

Peter Izsak, Moshe Berchansky, and Omer Levy. 2021. How to train BERT with an academic budget. In *Proceedings of the 2021 conference on empirical meth-*

ods in natural language processing, pages 10644–10652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jaap Jumelet, Michael Hanna, Marianne De Heer Kloots, Anna Langedijk, Charlotte Pouw, and Oskar van der Wal. 2023. ChapGTP, ILLC's attempt at raising a BabyLM: Improving data efficiency by automatic task formation. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Jean Kaddour. 2023. The minipile challenge for data-efficient language models. *CoRR*, abs/2304.08442.

Nikhil Kandpal, Brian Lester, Mohammed Muqeeth, Anisha Mascarenhas, Monty Evans, Vishal Baskaran, Tenghao Huang, Haokun Liu, and Colin Raffel. 2023. Git-theta: A git extension for collaborative development of machine learning models. 202:15708–15719.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

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.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, Paul Rottger, Tristan Thrush, and Scott Hale. 2022. Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1352–1368, Seattle, United States. Association for Computational Linguistics.

Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A. Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. In *First Workshop on Interpolation Regularizers and Beyond at NeurIPS 2022*.

Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. 2023. Stack more layers differently: High-rank training through low-rank updates. *CoRR*, abs/2307.05695.

Tal Linzen. 2019. What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, 95(1):e99–e108.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. 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*, 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. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery, and Lisa Beinborn. 2023. Climb – curriculum learning for infant-inspired model building. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Mark Mazumder, Colby R. Banbury, Xiaozhe Yao, Bojan Karlas, William Gaviria Rojas, Sudnya Frederick Diamos, Greg Diamos, Lynn He, Douwe Kiela, David Jurado, David Kanter, Rafael Mosquera, Juan Ciro, Lora Aroyo, Bilge Acun, Sabri Eyuboglu, Amirata Ghorbani, Emmett D. Goodman, Tariq Kane, Christine R. Kirkpatrick, Tzu-Sheng Kuo, Jonas Mueller, Tristan Thrush, Joaquin Vanschoren, Margaret Warren, Adina Williams, Serena Yeung, Newsha Ardalani, Praveen K. Paritosh, Ce Zhang, James Zou, Carole-Jean Wu, Cody Coleman, Andrew Y. Ng,

Peter Mattson, and Vijay Janapa Reddi. 2022. Dataperf: Benchmarks for data-centric AI development. *CoRR*, abs/2207.10062.

R. Thomas McCoy, Robert Frank, and Tal Linzen. 2020. Does syntax need to grow on trees? sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics*, 8:125–140.

Ian R. McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, Derik Kauffman, Aaron T. Kirtland, Zhengping Zhou, Yuhui Zhang, Sicong Huang, Daniel Wurgaft, Max Weiss, Alexis Ross, Gabriel Recchia, Alisa Liu, Jiacheng Liu, Tom Tseng, Tomasz Korbak, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*.

Maggie Mi. 2023. Mmi01 at the BabyLM challenge: Linguistically motivated curriculum learning for pretraining in low-resource settings. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *CoRR*, abs/2203.13112.

Omar Momen, David Arps, and Laura Kallmeyer. 2023. Increasing the performance of cognitively inspired data-efficient language models via implicit structure building. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. 2022. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1352–1368, Dublin, Ireland. Association for Computational Linguistics.

Aaron Mueller and Tal Linzen. 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11237–11252, Toronto, Canada. Association for Computational Linguistics.

Nikita Nangia and Samuel R. Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the GLUE benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575, Florence, Italy. Association for Computational Linguistics.

Toan Q. Nguyen and Julian Salazar. 2019. Transformers without tears: Improving the normalization of self-attention. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Miyu Oba, Akari Haga, Akiyo Fukatsu, and Yohei Oseki. 2023. CoNLL shared task BabyLM challenge: Curriculum learning based on sentence complexity approximating language acquisition. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Mattia Opper, J. Morrison, and N. Siddharth. 2023. On the effect of curriculum learning with developmental data for grammar acquisition. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Ludovica Pannitto and Aurélie Herbelot. 2020. Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 165–176, Online. Association for Computational Linguistics.

Isabel Papadimitriou and Dan Jurafsky. 2020. Learning Music Helps You Read: Using transfer to study linguistic structure in language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6829–6839, Online. Association for Computational Linguistics.

Andy Perfors, Joshua B. Tenenbaum, and Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulic, and Edoardo Maria Ponti. 2023. Modular deep learning. *CoRR*, abs/2302.11529.

Steven Piantadosi. 2023. Modern language models refute chomsky's approach to language. *Lingbuzz*. Preprint.

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

Jacob Portes, Alexander R. Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. 2023. MosaicBERT: How to train BERT with a lunch money budget. In *Workshop on Efficient Systems for Foundation Models at ICML2023*.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2021. DynaSent: A dynamic benchmark for sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 2388–2404, Online. Association for Computational Linguistics.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Shortformer: Better language modeling using shorter inputs. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5493–5505, Online. Association for Computational Linguistics.

Irina Proskurina, Guillaume Metzler, and Julien Velcin. 2023. Mini minds: Exploring Bebeshka and Zlata baby models. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Colin Raffel. 2023. Building machine learning models like open source software. *Communications of the ACM*, 66(2):38–40.

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.

Florencia Reali and Morten H. Christiansen. 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, 29(6):1007–1028.

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.

David Samuel. 2023. Mean BERTs make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

David Samuel, Andrey Kutuzov, Lilja Øvrelid, and Erik Velldal. 2023. Trained on 100 million words and still in shape: BERT meets British National Corpus. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1954–1974, Dubrovnik, Croatia. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Noam Shazeer. 2020. GLU variants improve transformer. *CoRR*, abs/2002.05202.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron Courville. 2021. StructFormer: Joint unsupervised induction of dependency and constituency structure from masked language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7196–7209, Online. Association for Computational Linguistics.

Sasha Sheng, Amanpreet Singh, Vedanuj Goswami, Jose Magana, Tristan Thrush, Wojciech Galuba, Devi Parikh, and Douwe Kiela. 2021. Human-adversarial visual question answering. In *Advances in Neural Information Processing Systems*, volume 34, pages 20346–20359.

Sam Shleifer, Jason Weston, and Myle Ott. 2021. Normformer: Improved transformer pretraining with extra normalization. *CoRR*, abs/2110.09456.

Massimo Stella, Nicole M. Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, 7(1):46730.

Julius Steuer, Marius Mosbach, and Dietrich Klakow. 2023. GPT-like models are bad babies: A closer look into the relationship of linguistic competence and psycholinguistic measures. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

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):339–374.

Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30.

Lukas Thoma, Ivonne Weyers, Erion Çano, Stefan Schweter, Jutta L. Mueller, and Benjamin Roth. 2023. Cogmemlm: Human-like memory mechanisms improve performance and cognitive plausibility of LLMs. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

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*. Association for Computational Linguistics (ACL).

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32.

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. 2022. *Artificial Neural Networks as Models of Human Language Acquisition*. PhD Thesis, New York University.

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.

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

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

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020c. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99, Online. Association for Computational Linguistics.

Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. 2023. Should you mask 15% in masked language modeling? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2985–3000, Dubrovnik, Croatia. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–44.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Lukas Wolf, Eghbal A. Hosseini, Greta Tuckute, Klemen Kotar, Alex Warstadt, Ethan Wilcox, and Tamar I Regev. 2023. WhisBERT: Multimodal text-audio language modeling on 100m words. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

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*. Association for Computational Linguistics (ACL).

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. Resolving interference when merging models. *CoRR*, abs/2306.01708.

Yahan Yang, Elior Sulem, Insup Lee, and Dan Roth. 2023. Penn & BGU BabyBERTa+ for strict-small BabyLM challenge. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Aditya Yedetore, Tal Linzen, Robert Frank, and R. Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9370–9393, Toronto, Canada. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. 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 pre-trained transformer language models. *CoRR*, abs/2205.01068.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Zheyu Zhang, Han Yang, Bolei Ma, David Rügamer, and Ercong Nie. 2023. Baby's CoThought: Leveraging large language models for enhanced reasoning in compact models. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Xingmeng Zhao, Tongnian Wang, Sheri Osborn, and Anthony Rios. 2023. BabyStories: Can reinforcement learning teach baby language models to write better stories? In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

Ömer Veysel Çağatan. 2023. ToddlerBERTa: Exploiting BabyBERTa for grammar learning and language understanding. In *Proceedings of the BabyLM Challenge*. Association for Computational Linguistics (ACL).

A Data Source Descriptions

CHILDES. The Child Language Data Exchange System (CHILDES; MacWhinney, 2000) is a multilingual database compiling transcriptions from numerous researchers of adult–child interactions in a range of environments, from structured laboratory activities to the home. Huebner and Willits (2021) further process CHILDES, selecting only interactions with American English-speaking children ages 0–6, removing all child utterances, and tokenizing the data. The resulting dataset[18] contains about 5M words.

British National Corpus. The BNC (Consortium, 2007) is a 100M word multi-domain corpus of British English from the second half of the 20$^{\text{th}}$ century. We select only the dialogue portion of the corpus, totaling about 10M words.

Children's Book Test. CBT is a compilation of over a hundred children's books from Project Gutenberg by Hill et al. (2016). The dataset was originally released with a set of questions for testing named entity prediction, which we do not include in the pretraining data.

Children's Stories Text Corpus. This dataset consists of manually selected children's stories from Project Gutenberg. It was compiled by Bensaid et al. (2021) for the development of a story generation system.

Project Gutenberg. The Standardized Project Gutenberg Corpus (Gerlach and Font-Clos, 2020) is a curated and preprocessed selection of over 50k literary books in the public domain from Project Gutenberg totaling over 3B tokens.[19] This distribution comes with extensive metadata that allows us to filter texts by language and date.

OpenSubtitles. This dataset (Lison and Tiedemann, 2016) is a compilation of publicly available subtitles from TV and movies on a third-party website.[20] We use only the English portion.

QED. The QCRI Educational Domain Corpus (formerly QCRI AMARA Corpus; Abdelali et al., 2014) consists of volunteer-written subtitles for educational videos. We use only the English portion.

Wikipedia. Wikipedia is a volunteer-authored encyclopedia hosted by the Wikimedia Foundation. We use only the English portion.

Simple English Wikipedia. Simple English is classified as a separate language in Wikipedia, thus the texts here are disjoint from those in English Wikipedia. The texts use shorter sentences and high-frequency vocabulary and avoid idioms.

Switchboard Corpus. The Switchboard Corpus (Godfrey et al., 1992) is a collection of transcribed telephone conversations between pairs of strangers. We accessed the text through the Switchboard Dialog Act Corpus (Stolcke et al., 2000).

[18] https://github.com/phueb/BabyBERTa/blob/master/data/corpora/aochildes.txt
[19] https://gutenberg.org/
[20] http://opensubtitles.org/

B Evaluation Data Details

As described in Section 5.2, we filter out evaluation examples that do not have lexical overlap with the *Strict-Small* pretraining corpus. Here, we present the number of training and test examples for each evaluation task after filtering. This allows us to partially control for the confound of the language style of most NLP tasks not aligning well with the pretraining corpus that we constructed. However, we only control for lexical content: other factors, such as sentence length, syntactic complexity, and overall linguistic style, remain distinct between our corpus and these tasks. In the future, it would be helpful for researchers to focus on designing tasks on which both children *and* language models can be reasonably evaluated.

Note, too, that our filtering procedure means that we cannot directly compare results obtained from the BabyLM Challenge to prior evaluations using the full datasets. We use a subset of the training and evaluation examples, and therefore can only compare between models evaluated on our version of these tasks.

	Task	|Train|	|Test|
BLiMP	Anaphor Agreement	–	1956
	Argument Structure	–	8248
	Binding	–	6738
	Control Raising	–	4526
	Determiner-Noun Agreement	–	7542
	Ellipsis	–	1732
	Filler-Gap	–	6426
	Irregular Forms	–	1965
	Island Effects	–	2676
	NPI Licensing	–	6586
	Quantifiers	–	3882
	Subject-Verb Agreement	–	5535
BLiMP Supplement	Hypernym	–	860
	Question-Answer Congruence (easy)	–	64
	Question-Answer Congruence (tricky)	–	165
	Subject-Auxiliary Inversion	–	4099
	Turn-taking	–	280
(Super)GLUE	CoLA	8164	1019
	SST-2	50528	508
	MRPC	1579	177
	QQP	243498	26889
	MNLI	259780	6562
	MNLI-mismatched	259780	6284
	QNLI	43917	2286
	RTE	858	99
	BoolQ	2072	723
	MultiRC	4637	913
	WSC	487	83
MSGS	Control Raising (Control)	6570	6731
	Lexical Content (Control)	9086	9100
	Main Verb (Control)	8166	8249
	Relative Position (Control)	9068	9046
	Syntactic Category (Control)	8930	8824
	Control Raising–Lexical Content	6816	6910
	Control Raising–Relative Token Position	8166	8167
	Main Verb–Lexical Content	7306	7378
	Main Verb–Relative Token Position	8177	8059
	Syntactic Category–Lexical Content	8181	7597
	Syntactic Category–Relative Position	9159	8298

Table 4: Number of training and test examples for each BabyLM evaluation task. We show the number of examples *after* filtering based on the pre-training corpus vocabulary (Section 5.2).

C Examples from the BLiMP Supplement

Contrast name	Acceptable sentence	Unacceptable sentence
BASE_AND_HYPONYM/ HYPERNYM	If he is growing herbs, then he is growing plants.	If he is growing herbs, then he is growing basil.
BASE_NEG_AND_HYPERNYM_NEG/ CONVERSE	If he isn't growing herbs, that means he isn't growing basil.	If he isn't growing basil, that means he isn't growing herbs.
BASE_NEG_AND_HYPERNYM_NEG/ HYPONYM_NEG	If he isn't growing herbs, that means he isn't growing basil.	If he isn't growing herbs, that means he isn't growing plants.
HYPERNYM_AND_BASE/ CONVERSE	If he is growing basil, that means he is growing herbs.	If he is growing herbs, that means he is growing basil.
HYPERNYM_AND_BASE/ OTHER	If he is growing basil, then he is growing herbs	If he is growing basil, then he is growing flowers.
HYPERNYM_AND_OTHER_NEG/ BASE_NEG	He is growing basil, therefore he isn't growing flowers.	He is growing basil, therefore he isn't growing herbs.

Table 5: Representative examples from the HYPERNYMS test suite of the BLiMP supplement.

Type	Length	Acceptable dialogue	Unacceptable dialogue
single	short	David: Should you quit?\n Sarah: No, I shouldn't.	David: Should she quit?\n Sarah: No, I shouldn't.
single	long	Did they try to finish it on time or not?\n No, they didn't.	Did we try to finish it on time or not?\n No, they didn't.
double	short	A: Did we say that you finished?\n B: Yes, you did.	A: Did you say that you finished?\n B: Yes, you did.
double	long	"Did you say that you will go somewhere after the movie is over?" he asked.\n "No, I didn't," she said.	"Did you say that you will go somewhere after the movie is over?" he asked.\n "No, you didn't," she said.

Table 6: Representative examples from the TURN-TAKING test suite of the BLiMP supplement.

Contrast name	Dif.	Acceptable dialogue	Unacceptable dialogue
ANIMATE VS. INANIMATE	easy	A: What did you purchase?\n B: Bread.	A: What did you purchase?\n B: David.
INANIMATE VS. ANIMATE	easy	"Who played the piano?" he asked. "A teacher played the piano," she said.	"Who played the piano?" he asked. "A car played the piano," she said.
LOC VS. NP	easy	David: Where did you put it?\n Sarah: Behind the sofa.	David: Where did you put it?\n Sarah: Eggs.
ANIMATE VS. INANIMATE	tricky	David: Who mopped?\n Sarah: A doctor.	David: Who mopped?\n Sarah: The tiles.
LOC VS. NP	tricky	A: Where were you reading?\n B: By the lake.	A: Where were you reading?\n B: An essay.
TEMP VS. NP	tricky	When did you eat?\n Several minutes ago.	When did you eat?\n Dinner.
EXPL VS. NP	tricky	"Why were you reading?" he asked. "For fun," she said.	"Why were you reading?" he asked. "A book," she said.
NUM VS. NP	tricky	A: How many do you teach?\n B: A few.	A: How many do you teach?\n B: History.
MANNER VS. NP	tricky	David: How did you vacuum?\n Sarah: I vacuumed quickly.	David: How did you vacuum?\n Sarah: I vacuumed the patio.

Table 7: Representative examples from the QUESTION-ANSWER CONGRUENCE test suite of the BLiMP supplement.

D Subtask Results

Here, we present a more detailed breakdown of results by subtask. Each task has a subsection containing a table where results are described, as well as a textual description containing and overview of the main takeaways for each task.

D.1 MSGS

Matthews correlation coefficients on MSGS (Table 8) were largely negative, indicating that language models trained at this scale tend to prefer surface features over linguistic features in ambiguous contexts. However, certain models demonstrated a much stronger preference for linguistic features in specific contexts: ELC-BERT showed high positive scores on average (sometimes significantly higher than Llama 2), as did Contextualizer. This shows us that architectural modifications can significantly improve scores, as can princpled approaches to curriculum learning.

In general, comparable models trained on the *Strict* corpus have higher MCCs than those trained on the *Strict-Small* corpus, but not always. This suggests that, while more pretraining data generally lead to stronger syntactic inductive biases, these preferences may depend on the features being compared, and that this will not always be the case depending on the architecture used.

	Model	Macro average	Ctl–Raising / Lexical content	Ctl–Raising / Relative position	Main verb / Lexical content	Main verb / Relative position	Syntactic cat. / Lexical content	Syntactic cat. / Relative position
	Llama 2 (Touvron et al., 2023)	-0.24	**0.93**	0.23	-0.77	-0.96	-0.19	-0.74
	RoBERTa-base (Liu et al., 2019)	-0.37	0.46	-0.58	-0.95	-0.94	0.36	-0.57
Strict	ELC-BERT (Charpentier and Samuel, 2023)	0.10	-0.51	-0.46	0.71	**0.97**	**0.46**	-0.53
Strict	Boot-BERT (Samuel, 2023)	-0.22	0.37	-0.77	-0.99	0.96	-0.34	-0.58
Strict	McGill (Cheng et al., 2023)	-0.35	0.65	-0.70	-0.99	-0.73	0.17	-0.49
Strict	*Best Baseline (OPT-125M)*	-0.39	0.35	-0.70	-0.76	-0.99	0.34	-0.60
Strict-small	ELC-BERT (Charpentier and Samuel, 2023)	-0.01	0.02	-0.71	**0.95**	0.50	-0.26	-0.59
Strict-small	MLSM (Thoma et al., 2023)	-0.37	0.31	-0.56	-0.99	-0.49	-0.03	-0.44
Strict-small	McGill (Cheng et al., 2023)	-0.60	-0.68	-0.37	-1.00	-0.79	-0.35	-0.42
Strict-small	*Best Baseline (OPT-125M)*	-0.45	0.00	-0.70	-0.72	-0.77	0.13	-0.68
Loose	Contextualizer (Xiao et al., 2023)	**0.24**	**0.88**	**0.71**	-0.32	0.30	0.21	**-0.35**
Loose	McGill (Cheng et al., 2023)	-0.75	-0.56	-0.97	-0.99	-0.86	-0.66	-0.46
Loose	BabyStories (Zhao et al., 2023)	-0.71	-0.24	-0.99	-0.99	-0.99	-0.23	-0.78

Table 8: MSGS results for each ambiguous subtask for the top performing models (by overall score) from each track, as well as baselines and skylines. MCC (i.e., linguistic bias score) results presented, truncated to two decimal places.

D.2 BLiMP

Accuracies on BLiMP (Table **??**) show that bigger models do not, as a rule, perform better on targeted grammatical evaluation. RoBERTa is the best-performing skyline model, despite that Llama 2 has orders-of-magnitude more parameters and was trained on significantly more data. Among the BabyLM submissions, Boot-BERT generally performs best, with ELC-BERT and McGill's submission also performing well in general on the *Strict* and *Strict-Small* tracks. ELC-BERT and Boot-BERT are both based on LTG-BERT (Samuel et al., 2023), suggesting that this architecture is a good starting point for pretraining on developmentally plausible amounts of linguistic input.

Analyzing specific test suites, we see that unsurprisingly that models in all tracks typically perform best on agreement phenomena, though we find surprisingly high variability on ANAPHOR AGREEMENT. **?**

reported that ISLAND EFFECTS and QUANTIFIERS were the two most difficult test cases. We find that the best BabyLM submissions actually outperform Llama by a wide margin on ISLAND EFFECTS. However, QUANTIFIERS, on which most models achieve very consistent and mediocre results, is the one test suite on which the Llama 2 skyline is stronger.

D.3 BLiMP Supplement

Accuracies on the BLiMP supplement tasks (Table 9) demonstrate similar trends as those in the BLiMP tasks. As these individual test suites are new to this task, these fine-grained results are of particular interest. We find that the HYPERNYM test suite is clearly beyond the ability of language models. All models including the skylines perform very close to chance, suggesting either that their preferences are virtually random guessing, or they show systematic biases that essentially cancel out due to counterbalancing in the test data. However, we hesitate to conclude that these models have no knowledge of lexical entailment relations for two reasons: First, these test sentences are somewhat unnatural logical statements which are out-of-domain for the models, and second, there is less reason *a priori* to think that logically invalid statements have lower probability than valid statements.

Among the QUESTION–ANSWER CONGRUENCE test suites, we do indeed find that the "tricky" examples are far more difficult than the "easy" ones. The "tricky" set is highly discriminative, due probably to its adversarial nature, telling us that most models are easily fooled by locally coherent distractor answers and pay too little attention to cross-sentential long-distance dependency between a *wh*-word and a congruent answer. Only the top-performing models in the *Strict* track score better than chance, and the RoBERTa skyline outperforms all models by a wide margin.

The tests for SUBJECT–AUXILIARY INVERSION are relatively easy, with the best models reaching near-perfect accuracy. TURN TAKING is highly discriminative, with some models performing at or near chance, while the best model achieves accuracy over 90%. Again, ELC-BERT outperforms the skylines. This may be due in part to the fact that transcribed dialogue is a relatively large proportion of the BabyLM training data, compared to the training data for typical pretrained language models.

	Model	Macro average	Hypernym	Q–A congruence (easy)	Q–A congruence (tricky)	Subject–aux inversion	Turn taking
	Llama 2	0.74	**0.50**	0.85	0.63	0.91	0.83
	RoBERTa	0.75	0.48	**0.87**	**0.72**	**0.98**	0.73
Strict	ELC-BERT	**0.76**	0.47	**0.85**	**0.63**	0.94	**0.92**
Strict	Boot-BERT	0.72	0.45	0.75	0.58	**0.96**	0.86
Strict	McGill	0.71	0.46	0.84	0.58	0.82	0.83
Strict	OPT	0.67	0.46	0.76	0.47	0.85	0.82
Strict-small	ELC-BERT	0.67	0.48	0.68	0.44	0.88	0.83
Strict-small	MLSM	0.57	0.47	0.70	0.33	0.82	0.52
Strict-small	McGill	0.58	0.49	0.73	0.35	0.77	0.57
Strict-small	OPT	0.52	**0.50**	0.54	0.31	0.70	0.57
Loose	Contextualizer	0.63	0.47	0.73	0.42	0.91	0.62
Loose	McGill	0.56	0.49	0.64	0.29	0.80	0.61
Loose	BabyStories	0.64	0.49	0.71	0.50	0.79	0.73

Table 9: BLiMP Supplement accuracies for each subtask for the top performing systems (by overall score), best baseline, and skylines. For each subtask, we mark the best performing system for each track, and the **best** non-skyline and **best** performing system overall.

D.4 GLUE/SuperGLUE

Scores on (Super)GLUE tasks (Table 10) show that ELC-BERT is generally the best-performing system in both the *Strict* and *Strict-Small* tracks, and that Boot-BERT is also highly effective in the *Strict* track. Contextualizer also performs well. This largely confirms findings from the BLiMP and BLiMP Supplement tasks: LTG-BERT is an effective architecture for pretraining on smaller corpora, and curriculum learning can improve performance over a naïve corpus ordering.

	Model	Macro average	CoLA	SST-2	MRPC	QQP	MNLI	MNLI-mm	QNLI	RTE	BoolQ	MultiRC	WSC
	Llama 2	**0.83**	**0.63**	**0.95**	0.87	0.81	0.85	**0.87**	0.89	**0.81**	**0.85**	**0.86**	**0.75**
	RoBERTa	0.78	0.62	0.93	0.88	0.87	**0.86**	0.85	**0.92**	0.61	0.76	0.68	0.61
Strict	ELC-BERT	**0.78**	**0.59**	**0.92**	**0.90**	**0.88**	0.84	0.83	0.89	0.64	**0.73**	0.72	**0.62**
Strict	Boot-BERT	**0.78**	0.57	**0.92**	0.89	**0.88**	**0.85**	**0.84**	**0.91**	**0.65**	0.72	**0.73**	0.61
Strict	McGill	0.72	0.49	0.89	0.83	0.86	0.79	0.79	0.84	0.53	0.66	0.65	0.61
Strict	OPT	0.70	0.36	0.88	0.82	0.83	0.76	0.77	0.83	0.63	0.66	0.60	0.54
Strict-small	ELC-BERT	0.73	0.47	0.86	0.87	0.86	0.78	0.79	0.84	0.60	0.69	0.68	**0.62**
Strict-small	MLSM	0.70	0.41	0.90	0.78	0.85	0.75	0.76	0.82	0.59	0.66	0.58	0.61
Strict-small	McGill	0.69	0.41	0.87	0.79	0.81	0.73	0.74	0.79	0.54	0.66	0.62	0.61
Strict-small	OPT	0.62	0.15	0.84	0.74	0.78	0.67	0.69	0.65	0.55	0.65	0.51	0.59
Loose	Contextualizer	0.72	0.56	0.90	0.83	0.85	0.77	0.78	0.83	0.53	0.68	0.64	0.59
Loose	McGill	0.68	0.37	0.88	0.77	0.83	0.73	0.75	0.78	0.49	0.67	0.60	0.61
Loose	BabyStories	0.60	0.00	0.84	0.82	0.66	0.59	0.64	0.79	0.53	0.67	0.46	0.61

Table 10: (Super)GLUE results for each subtask for the top performing systems (by overall score), best baseline, and skylines. For each subtask, we mark the best performing system for each track, and the **best** non-skyline and **best** performing system overall.

E Age of Acquisition Prediction Results

Here, we present scores, separated by track, for each model that evaluated on the age of acquisition (AoA) prediction task (Table 11). We also compare to the best-performing baseline within each track, as in Table 2.

Almost all submissions which evaluated on the AoA prediction task were in the *Strict-Small* track. Here, no model achieved closer predictions than the OPT-125M baseline, though many got very close. In the *Strict* track, BabyStories achieved very close scores to the OPT-125M baseline.

	Model	Mean average deviation ↓			
		Overall	Nouns	Predicates	Function Words
Strict	BabyStories (GPT2-Large-PPO) (Zhao et al., 2023)	2.05	1.98	**1.82**	2.63
Strict	*Best Baseline (OPT-125M)*	**2.04**	**1.97**	1.83	**2.61**
Strict-Small	GPT-Wee (16k (cu.)) (Bunzeck and Zarrieß, 2023)	2.06	2.00	1.83	2.58
Strict-Small	Bebeshka (Proskurina et al., 2023)	2.06	1.98	1.84	2.66
Strict-Small	Zlata (Proskurina et al., 2023)	2.07	1.99	1.83	2.67
Strict-Small	Too Much Information (Edman and Bylinina, 2023)	2.05	1.99	1.85	2.58
Strict-Small	Mmi01 (RARITY) (Mi, 2023)	2.05	**1.97**	1.85	2.64
Strict-Small	Baby Llama (Timiryasov and Tastet, 2023)	2.06	1.99	1.84	2.63
Strict-Small	Lil-Bevo-X (Govindarajan et al., 2023)	2.05	1.99	1.85	2.59
Strict-Small	*Best Baseline (OPT-125M)*	**2.03**	1.98	**1.81**	**2.57**

Table 11: Mean average deviation (MAD) in months across cross-validation folds when predicting the age of acquisition of words. Lower MAD scores are better. We present all systems that evaluated on AoA prediction, as well as the baseline model with the best scores per track. We **bold** the highest-scoring system for each task within each track.

F Summary of Each Submission

GPT-wee (Bunzeck and Zarrieß, 2023). This paper tests various approaches to reordering the examples based on word and sentence statistics. The motivation comes from usage-based linguistics and the idea that frequent lexical items, such as phrases or common groups of words, are learned early (rather than words, for instance). They also find that training more—up to 10 epochs—helps, and that a medium-sized model might be as good as larger models.

Tiny Language Models with Multiplex Networks (Fields et al., 2023). This approach leverages multimodal data (including text/visual data and sensorimotor data) as part of the embeddings to an ELECTRA language model. The proposed models are very small (as few as 7M parameters) and perform well on BLiMP. For reference, the baseline models contain 125M to 220M parameters.

Mini Minds (Proskurina et al., 2023). This submission explores how scaling down models (in terms of number of parameters) can help in low-data settings. The authors conduct a parameter search for scaled-down versions of GPT-2 and RoBERTa, and find that optimal models have around a 2-to-1 ratio of attention heads to layers. They train two models and find that they perform about as well as larger parameter count models on GLUE. Furthermore, the authors test their models on an ethical reasoning benchmark and find that the small models perform about as well as models which have about ten times the parameters.

Grammar induction pretraining (Chen and Portelance, 2023). This submission introduces syntactic bias into the static token embeddings of an LM. An unsupervised grammar induction system is trained on a 1-million word subset of the *Strict-Small* corpus, and the resulting static token embeddings are used to initialize the LM token embeddings. Although the results improve over the BabyLM *Strict-Small* baseline, similar improvements are observed with a custom baseline model using randomly initialized token embeddings. Thus, there is no evidence that the grammar induction step had a positive impact on LM results.

ChapGTP (Jumelet et al., 2023). This work explores how targeted data augmentation can improve the performance of masked language models in the *Strict-Small* track. The authors used regex patterns to extract common phrases from the GLUE tasks and then used these patterns to generate follow-up questions that served as additional training data. They also found that increasing the training epochs up to 200 epochs continues to help performance.

BabyBerta+ (Yang et al., 2023). The submission replicates the BabyBERTa training setup (Huebner et al., 2021) and tests its ability after pretraining on the *Strict-Small* corpus. They find that a small model trained on many epochs keeps improving and becomes better than baseline models in grammatical aspects, but not downstream tasks.

Keeping Training Simple for BabyLMs (Edman and Bylinina, 2023). This paper proposes a variety of complexity metrics for reordering the BabyLM *Strict-Small* data from simple to complex. Compared to no curricula and reversed curricula, the proposed curricula do not result in consistent performance improvements on the BabyLM evaluation tasks. However, reducing the context length to 32 (from the baselines' 128) results in significant and consistent performance improvements.

Can Training Neural Language Models on a Curriculum with Developmentally Plausible Data Improve Alignment with Human Reading Behavior? (Chobey et al., 2023). This paper explores surprisal-based curricula for pretraining on the *Strict-Small* dataset of the BabyLM challenge. The authors use an ensemble of LSTM "teacher" models to rank sentences by average surprisal, on which a final OPT

model is trained. Results are mixed. The authors find that their model does not outperform a random baseline. However, when this model is further trained on the randomly-ordered training dataset after training on the curriculum-ordered data, it does beat the baseline. As an additional analysis, the authors investigate the ability of their model to predict human reading times for syntactically complex sentences, finding that the model is not particularly good at the task, but that it is about equivalent to baselines which are trained on much larger datasets.

CLIMB (Martinez et al., 2023). This submission presents a thorough comparison of different approaches to curriculum learning in the *Strict-Small* setting. They consider three main criteria for curriculum design: the size of the input vocabulary, the difficulty of the training sample, and the size of the output space for MLM prediction. They conduct experiments exploring eight different curricula sorted into these three main approaches. While there are many small differences in performance among these settings, curricula provide no consistent improvements over more naive training algorithms.

Acquiring Linguistic Knowledge from Multimodal Input (Amariucai and Warstadt, 2023). The authors explored whether vision-language co-training helps the learning of linguistic knowledge. They trained models on Wiki texts with images using the state-of-the-art multi-modality model (FLAVA). After varying the amount of training data and how many images are used, the authors found that visual input only provides a slight improvement on grammar benchmarks for 10M-word training, but not for 100M-word training.

GPT-like Models are Bad Babies (Steuer et al., 2023). This paper trains a decoder-only model, trying different hyperparameters, including reordering the training data by different orders (based on cues which did not improve over regular shuffling), different sizes, layer widths, among other features. The main focus of the paper is to test if models that perform better on BabyLM evaluation tasks are also better at modeling reading difficulty in humans. Surprisingly, models performing better on BabyLM tasks performed *less* well in modeling reading difficulty.

Baby's CoThought (Zhang et al., 2023). This system leverages a large language model, GPT-3.5-Turbo, to reformat semantically unrelated sentences into cohesive paragraphs. In low-data settings, this approach can form better training examples for language models; the proposed approach results in improvements across BLiMP tasks, though performance is not significantly different on (Super)GLUE or MSGS. Note that the LLM is trained on far more than 100M words, so this submission technically does not qualify under any track. However, this method does improve the sample efficiency of the student model, and it aids our understanding of what types of data are best for supervising smaller language models.

ToddlerBERTa (Çağatan, 2023). This paper conducts a thorough hyperparameter investigation of the BabyBERTa model, exploring different options for model sizes and training algorithms. The author finds that larger models tend to perform better.

CogMemLM (Thoma et al., 2023). This work explores an approach to word segmentation and tokenization that is intended to model vocabulary growth during learning. A vocabulary is cumulatively built using a cognitively-inspired model of word segmentation, in which strings are split into chunks based on an activation weight which changes throughout training depending on how often the chunk is observed together. While the approach achieves consistent improvements over the BabyLM *Strict* baseline results, it is not clear whether these improvements are due to the segmentation scheme or other hyperparameter modifications.

BabyStories (Zhao et al., 2023). This paper investigates how reinforcement learning from human feedback (RLHF) improves the performance of causal language models pretrained on small scales of datasets. The authors report that models finetuned by RLHF on short stories yield better performance on language understanding benchmarks, though this improvement is only observed on larger models. Their findings suggest that benefiting from RLHF requires a large number of trainable parameters.

Byte-ranked Curriculum Learning (DeBenedetto, 2023). This paper proposes a curriculum learning approach for reordering data based on non-linguistic metrics. Specifically, they choose the order in which

datasets are shown to the model starting from the minimal amount of bytes per sentence and going up. This happens to also start from spoken data and follow with text data later. The paper also shows that a larger model as well as more epochs improves the results.

McGill BabyLM Submission (Cheng et al., 2023). This paper finds that changes to the data format have large positive impacts. Specifically, not using sequence packing, using sentences and not documents as examples, not truncating, and reducing maximum sequence length are each highly effective. By contrast, adding supervision from POS tags and using unsupervised syntactic induction have negligible impact.

Mean BERTS make erratic language teachers (Samuel, 2023). This submission presents Boot-BERT, a latent bootstrapping approach to language modeling in low resource settings. In the latent bootstrapping set-up, a student model is trained to produce predictions over words as well as to match contextualized embeddings from a teacher model. In turn, the teacher's embeddings are obtained via a moving average of the student's. The authors use LTG-BERT (Samuel et al., 2023) as an encoder backbone, as well as for a baseline.[21] They find that their Boot-BERT outperforms LTG-BERT for some of the BabyLM tasks, including GLUE for both the *Strict* and *Strict-Small* tracks.

Every Layer Counts BERT (ELC-BERT) (Charpentier and Samuel, 2023). This submission takes as its starting point the very effective LTG-BERT architecture from Samuel et al. (2023) and modifies it such that the input to each layer is a weighted sum of the outputs of all previous layers, where the weights can be learned but also biased by initialization. Several variations are explored, including equal initial weights, and initial weights biased towards the previous layer. Results on BabyLM evaluations do not strongly suggest that any one variant is clearly better than the LTG-BERT baseline, though all models perform significantly better than the BabyLM RoBERTa baseline. Additionally, inspection of the learned weights for combining previous layer outputs suggests that the most important outputs are from the previous few layers and the static embedding layer.

WhisBERT (Wolf et al., 2023). In this submission, the authors explore whether text-and-audio co-training helps model performance on BLiMP tasks. After pretraining a multi-modal model (FLAVA) on 100M words with or without their corresponding word-aligned speech, they find that the speech-augmented model outperforms the text-only model on 11 out of 17 grammatical tasks.

Surprisal-based active curriculum learning (Hong et al., 2023). This submission combines curriculum and active learning to schedule training order for models. The authors use n-gram surprisals to determine the sentences with the highest surprisal and then train their models on structurally similar examples to these high-surprisal sentences. Models with active curriculum learning show noticeable performance gains in (Super)GLUE but underperform the models without such learning on MSGS.

Linguistically Motivated Curriculum Learning (Mi, 2023). This submission tests 6 linguistic metrics of complexity as curriculum learning approaches. On the *Strict-Small* track, this approach succeeds in finding improvements over training on the whole corpus in a random order.

Baby Llama (Timiryasov and Tastet, 2023). This submission proposes a knowledge distillation approach with two teacher models (a 300M-parameter Llama model and 700M-parameter GPT-2 model) trained on the *Strict-Small* corpus. These are distilled into a 58M-parameter Llama model called Baby Llama. The proposed model outperforms the BabyLM baselines, the teacher LMs, and a 58M-parameter Llama model trained from scratch on the *Strict-Small* data without distillation.

Curriculum learning based on sentence complexity approximating language acquisition (Oba et al., 2023). This submission assesses the impact of curriculum learning based on sentence complexity within the context of the *Strict-Small* task. The authors order training data based on three sentence-level complexity metrics: number of tokens, number of constituents, and max depth of the sentences'

[21] As described in §7.2, LTG-BERT makes multiple modifications to the standard Transformer encoder architecture: additional layer normalization (Shleifer et al., 2021), GEGLU feed-forward modules (Shazeer, 2020), disentangled attention following DeBERTa (He et al., 2021), and scaled weight initialization following (Nguyen and Salazar, 2019).

dependency parse. They find that the dependency-based ranking leads to better models, however, all curriculum-based models underperform a random baseline.

Masked Latent Semantic Modeling (Berend, 2023b). This paper adopts a method from Berend (2023a) called Masked Latent Semantic Modeling (MLSM) in which the target output distribution can be transformed from a one-hot distribution over the vocabulary into a sparse distribution over latent "semantic property" vectors. Then, the same kind of student-teacher optimization as in knowledge distillation is applied using this modified output distribution instead of the full vocabulary. MLSM on its own is found to lead to degradation in BLiMP performance, although combining MLSM with typical MLM training in a multitask setting leads to similar performance as MLM training alone.

Lil-Bevo (Govindarajan et al., 2023). This paper offered submissions to both *Strict-Small* and *Strict* tracks and used three design choices for LM training: (i) initially pretraining on music data, following work on transfer learning (Papadimitriou and Jurafsky, 2020), which suggested that musical structure may form a reasonable basis upon which to learn language structure; (ii) subsequently using a training curriculum starting from shorter sequences (128) before moving to longer ones (512), following insights from Press et al. (2021), and (iii) masking critical tokens necessary to perform some of the BLiMP subtasks (e.g., masking "not" for NPI-licensing). Taking final results into consideration alongside ablations, this team found that sequence length matters, music pretraining may help a little, and targeted MLM training seems to help (but only for some BLiMP subtasks, including NPI licensing and Argument Structure).

Contextualizer (Xiao et al., 2023). This paper sorts the corpora in the training dataset loosely based on their age of acquisition and reading difficulty. The authors then introduce techniques to begin and end the training with padding-separated datasets sorted from easy to hard, while the middle of the training employs a noisier padding and sorting strategy to improve the model's robustness. The final model performs similarly to its counterpart pretrained with thousands of times more data.

Implicit Structure Building (Momen et al., 2023). This submission introduces an unsupervised hierarchical bias into the transformer. The approach shows that such structural bias with StructFormer improves over the classic MLM Transformer approach. Improvements are not consistent across scenarios: the model excels in single-sentence or syntactic evaluation tasks, but less so in semantic tasks with multi-sentence inputs.

Pretraining LLMs using human-like development data (Bhardwaj et al., 2023). This submission trains RoBERTa, DistilBERT, and GPT-2 models on the *Strict* and *Strict-Small* data. They find that training DistilBERT for 60 epochs is better than 20 epochs. They also claim that the performance of the baseline RoBERTa model may not be replicable across random initializations and that hyperparameter searches should be more thorough to hedge against such outlier models.

On the Effect of Curriculum Learning with Developmental Data for Grammar Acquisition (Opper et al., 2023). This submission explores the effect of curriculum learning, using BabyBERTa models, on the *Strict-Small* data track. The authors contrast three types of curriculum learning: one that orders input by word frequency; one by sequence entropy; and one by increasing context length. They find that neither of these methods produces results above a baseline random presentation. In a series of follow-up experiments, the authors verify that model performance is linked to the amount of exposure to transcribed speech data and suggest that speech data is a good foundation for curriculum learning.

Difficulty-based Sentence Reordering (Borazjanizadeh, 2023). This study explores two broad approaches to dataset preprocessing to improve LM training in the 10M-word setting: data reordering (curriculum learning) and data cleaning. Results show that reordering a subset of the data by sentence difficulty may lead to marginal improvements, as long the local coherence of the samples is not damaged too greatly. However, the clearest improvements come from cleaning the data of incoherent, ungrammatical, or non-linguistic strings.

G Results Broken Down by GLUE / BLiMP Subtask

Figure 7: **Submission Results by GLUE subtask:** Points show the performance of each submission. Gray bars show the across-submission average in each category.

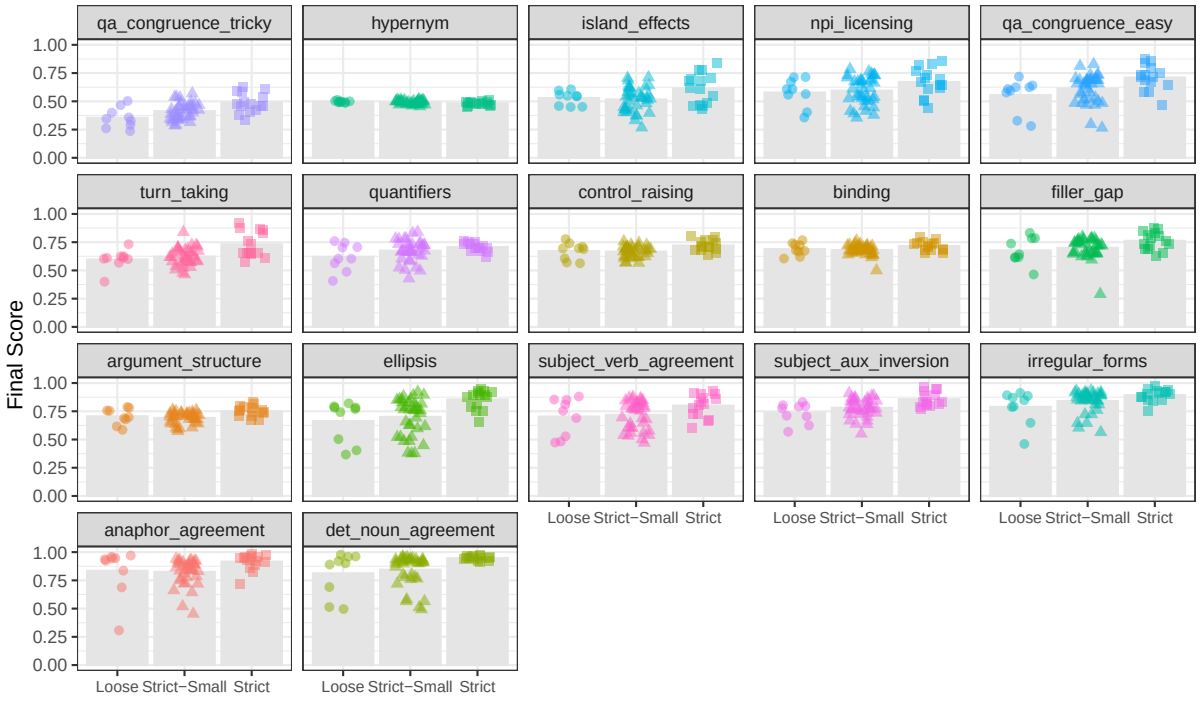

Figure 8: **Submission Results by BLiMP subtask:** Points show the performance of each submission. Gray bars show the across-submission average in each category.