Journal of Memory and Language 144 (2025) 104650

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Bigger is not always better: The importance of human-scale language modeling for psycholinguistics

Ethan Gotlieb Wilcox[f,*], Michael Y. Hu[b], Aaron Mueller[c], Alex Warstadt[a], Leshem Choshen[d], Chengxu Zhuang[d], Adina Williams[e,1], Ryan Cotterell[a,1], Tal Linzen[b,1]

[a] Department of Computer Science, ETH Zürich, Universitätstrasse 6, 8092, Zürich, Switzerland
[b] Department of Linguistics and Center for Data Science, New York University, 10 Washington Place, NewYork, NY, 10003, USA
[c] Khoury College of Computer Sciences, Northeastern University, 440 Huntington Avenue, Boston, MA, 02115, USA
[d] Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA, 02139, USA
[e] FAIR Laboratories, Meta Platforms Inc., 390 9th Ave, NewYork, NY, 10001, USA
[f] Department of Linguistics, Georgetown University, 1421 37th St NW, Washington DC, 20007, USA

ARTICLE INFO

Keywords:
Language modeling
Scaling
Cognitive modeling
Connectionist networks
Psycholinguistics
Language acquisition

ABSTRACT

When trained to place high probability on a training corpus, neural network language models can learn a surprising amount about language. Recent work has demonstrated that large performance improvements can arise from simply increasing, i.e., scaling, the size of the corpora they are trained on and the number of parameters in those models. Accordingly, many contemporary systems are trained on trillions of words. While largely beneficial to performance on language applications, scaling has several downsides for both computational psycholinguistics and natural language processing research. We discuss the scientific challenges presented by the scaling paradigm, as well as the benefits that would result from language models that can learn from human-scale data. In the second half of this paper, we report on findings from a recent effort to bring about human-scale language model pretraining: the first iteration of the BabyLM Challenge, a shared task organized by the authors that invited participants to train a language model on 100 million words or less. The challenge produced several concrete best practices for practitioners interested in small-scale language modeling. For cognitive scientists, the challenge demonstrated that robust linguistic generalizations can be learned by models trained on a human-scale dataset, though this is not yet achieved through cognitively plausible mechanisms. Furthermore, it established a population of "BabyLMs" that are all effective at data-efficient language learning. Studying such models can help us identify hypotheses for the computational mechanisms that underlie human language acquisition.

Introduction

Connectionist modeling, i.e., modeling based on neural networks, has been a core theoretical and empirical tool for psycholinguistics research over the past four decades (Christiansen & Chater, 1999; Elman, 1990; Smolensky, 1988). This approach, which provides a paradigm for explaining how processing emerges from learning and how symbolic structure can be implemented by distributed representations, has seen a recent resurgence due to the emergence of highly effective **language models** (LMs) based on neural networks. Language models are trained to fit the empirical distribution of words in a training corpus, and, through this process, learn a considerable amount about grammar and other aspects of language (Linzen & Baroni, 2021). Much of the practical success of LMs in applications has been driven by **scaling** language models (Hoffmann et al., 2022; Kaplan et al., 2020), i.e., increasing the number of model parameters, training them on larger and larger amounts of data, or often both. This article takes a critical look, from a psycholinguistics perspective, at scaling both the number of parameters and the size of the training corpus as a path to improve language models. In short, we argue that bigger is not always better, and that future success in connectionist modeling of psycholinguistic processes will require, alongside the current scaling paradigm, models that can be trained on corpora of a humanlike scale.

First, we discuss the impact of scaling on the relationship between connectionist modeling and linguistics, a relationship that has been characterized as "frictional" (Pater, 2019). We outline two key ways connectionist modeling can contribute to linguistics research, and argue

[*] Corresponding author.
E-mail address: ethan.wilcox@georgetown.edu (E.G. Wilcox).
[1] Senior authors.

https://doi.org/10.1016/j.jml.2025.104650
Received 15 November 2023; Received in revised form 28 April 2025; Accepted 29 April 2025

that the utility of language models for psycholinguistics research is limited by the current trend toward larger and more data-intensive models. We propose that a simple way to mitigate these issues is to devote effort toward building more data-efficient connectionist models trained on more developmentally plausible datasets in terms of size, genre, and modality.

Second, we discuss the impact of scaling on machine learning (ML) and natural language processing (NLP) research. Although generally beneficial for NLP applications, we argue that the scaling trend is not without its downsides. We highlight three specific issues with this paradigm. First, while the focus on evaluating language models based solely on performance incentivizes scaling and disincentivizes research into data-efficient models (Linzen, 2020), despite the fact that data-efficient models are essential when training data is scarce, for example, in low-resource languages or specialized domains. Second, smaller datasets are easier to curate and control for quality. Third, due to the cost of training models at scale, the focus on scale produces a high barrier to entry and an environment in which research teams might be relatively risk-averse. Both of these factors can potentially lead to scientific stagnation. Again, we propose that downsides can be mitigated by devoting efforts toward building and training models at smaller data scales. Such models can be prototyped and tested quickly and cheaply, allowing for broader participation and faster innovation in machine learning research.

In the remainder of the paper, we discuss a recent effort we led that was motivated by these concerns: the BabyLM Challenge (Warstadt et al., 2023), a shared task that invites participants to train language models on the amount of data available to a typical human language learner. The BabyLM Challenge was held at a large natural language processing conference in the fall of 2023 and received a large number of participants, as well as national press coverage (Whang, 2023). We identify several key technical findings from the challenge and discuss their implications for psycholinguistics. For NLP practitioners, we recommend one model architecture — called LTG-BERT — as a good starting point for small-scale language modeling; in Appendix A, we present several follow-up studies that investigate various features of this architecture, which was the winner of the challenge. We discuss two main takeaways from the BabyLM Challenge for cognitive scientists and psycholinguists. First, the challenge demonstrated that robust syntactic and semantic generalizations can be learned by neural language models trained on human-scale corpora, even though these training methods are often not cognitively plausible. Indeed, some of our best-performing models were just a few percentage points shy of human performance on grammatical acceptability tasks. Second, the challenge established a population of models that are all effective data-efficient language learners. Studying these "BabyLMs" can help us identify hypotheses for the computational mechanisms that underlie human language learning. We note, however, that all models were trained on English text and the extent to which these findings generalize across typologically diverse languages therefore remains an open question that should be ddressed in future research.

Scaling neural language models

Up until the mid-2010s, the dominant paradigm in natural language processing (NLP) was to build systems that combined a series of highly articulated, domain-specific components. For example, in a machine translation system of that epoch, one component would be responsible for aligning words between the source and target sentence, another component responsible for homonym disambiguation, and another for scoring the naturalness of the proposed text (Block, 1962; Brown, Della Pietra, Della Pietra, & Mercer, 1993). However, over the course of the last decade, this paradigm has changed. Nowadays, the best-performing NLP tools typically consist solely of a neural network LM optimized to predict the probability of a unit of text (or *token*) given its context.[2] After this initial training stage, referred to as **pretraining**, the LM can be adapted to particular applications, through training on a secondary objective, e.g., to predict the sentiment of a sentence. This second stage of training is called **fine-tuning**. In contrast to the classic NLP architecture that consists of a series of modular components, the neural paradigm involves just one single system, the LM, which can be adapted to a variety of tasks (Devlin et al., 2019). More recently, many contemporary LMs are adapted to new tasks by simply conditioning the model's output on several examples of the task, a technique known as **in-context learning** (Brown et al., 2020). Whereas previously, an engineer seeking to improve an NLP system could, in principle, improve one of its modular components in isolation and then reinsert it into the system, neural network language models often cannot be modularized in this fashion. Indeed, a priori, it is not clear what constitutes a module in a neural LM. Thus, improving the performance of such a holistic system often requires retraining the neural LM in its entirety.

In engineering better and better neural language models, the field of NLP has benefited from a larger trend in computer science: the growing amount of data and computing power afforded by modern computers (Coffman & Odlyzko, 2002; Schaller, 1997). Neural network–based systems that were capable of learning from larger amounts of data tended to outperform their competitors, even without architectural changes compared to those competitors. For example, one popular LM model, GPT-2 (Radford, Wu, Child, Luan, Amodei et al., 2019) performed substantially better on NLP tasks than its predecessor, GPT (Radford, Narasimhan, Salimans, Sutskever et al., 2018), even though the two have very similar architectures: the main difference between them is that GPT-2 has more parameters and was trained on more data. More broadly, one reason why models using the Transformer architecture (Vaswani et al., 2017), upon which GPT and GPT-2 are based, are so successful is that they were designed to enable computationally efficient training, which allows them to be trained on larger corpora. Repeated experiments all pointed towards the benefit of scaling, not only in natural language processing but in other domains such as vision, leading to the "bitter lesson" (Sutton, 2019): namely, that the best learning methods are general-purpose methods that can leverage the most data and compute.

But this observation — roughly, that bigger is better — raises several important questions, particularly because "bigger" is underspecified. When training neural language mdoels, three high-level elements need to be balanced: the size of the model, i.e., the number of trainable parameters, which we will refer to as P; the size of the training data, which we will refer to as T; and the number of computations that are performed during training, sometimes referred to informally as "compute", and which we will refer to as C.[3] Using this notation, scaling refers to the practice of balancing P, T, and C to optimize for the best model possible given one's budget. A growing body of work has explored this question in recent years, both for natural language technologies (Bahri, Dyer, Kaplan, Lee, & Sharma, 2024; Henighan et al., 2020; Hestness et al., 2017; Hoffmann et al., 2022; Muennighoff et al., 2024; Rosenfeld, Rosenfeld, Belinkov, & Shavit, 2019) as well as in other fields such as for computer vision (Zhai, Kolesnikov, Houlsby, & Beyer, 2022) or protein sequence models (Hesslow, Zanichelli, Notin, Poli, & Marks, 2022). Typically, because the compute is the limiting factor when training LLMs, scaling research seeks to uncover the optimal

[2] While in the original usage of the term the context used to predict the current token consisted only of the preceding ones in the sentence, this term has recently expanded to include models that have access to both the preceding and following context, sometimes referred to as *masked* language models, such as the BERT model presented in Devlin, Chang, Lee, and Toutanova (2019).

[3] Modern computers approximate real numbers with floating-point numbers. Compute costs are therefore measured most precisely in the number of floating-point operations or FLOPs. In this article, however, we will informally characterize compute costs as the number of times a model iterates over its training corpus during training.

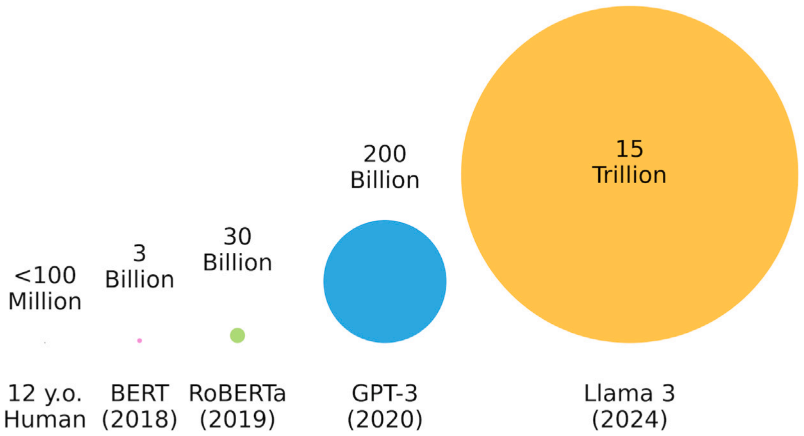


Fig. 1. Data Scale: Mainstream language models are trained on multiple orders of magnitude more word tokens than the amount available to a typical child.

choices for P and T for a fixed value of C. One influential approach proposed in Hoffmann et al. (2022) and Kaplan et al. (2020) is to empirically study this relationship for low values of P and T and then to extrapolate to higher values. Their results suggested that given a fixed compute budget, model architecture size, and training data size should scale proportionally, i.e., every time one doubles P, one should also double T. This has led to an ever-growing reliance on larger and larger model sizes and, importantly for our discussion, training corpora. State-of-the-art language models in 2025 are trained on trillions of words of text, and are referred to as **large language models** (LLMs) to distinguish them from their more modestly sized counterparts. For example, one recent system, Llama 3, was trained on as many as 15 trillion tokens (Dubey et al., 2024). The scale of the increasingly larger training corpora is visualized in Fig. 1, which compares the training corpus size of today's LMs with the typical amount of human linguistic experience at the onset of adolescence—under 100 million words for children growing up in the United States (Gilkerson et al., 2017).

The downsides of scaling for psycholinguistics

At the heart of this article lies a central question: How can neural language models further our scientific understanding of language? There is a certain irony in posing this question, given that connectionist architectures were not originally developed to process and manipulate text data, but rather to model human cognition (Elman, 1990; Rumelhart, McClelland, Group et al., 1986). Yet, because of their success in practical modeling applications, their cognitive-modeling origins have largely been overshadowed. In this section, we aim to bring neural networks back to these roots by illustrating two examples of how neural language models can advance research in linguistics and psycholinguistics (see also Linzen, 2019). We argue, however, that these contributions are only valid under certain conditions, which are often not met by large-scale models. Although our focus here is on the role of neural language models in the study of language, our arguments are broadly applicable to any neural-network architecture intended to model human cognition.

Stimulus–poverty arguments

The first type of contribution uses neural networks to assess stimulus-poverty arguments. Stimulus-poverty claims are used to argue for a particular perspective on how children learn language and have been influential in the linguistics literature since they were first introduced around fifty years ago (Chomsky, 1965, 1979). Stimulus-poverty arguments point out that the primary linguistic data available to children are compatible with a large number of hypotheses about how that data is underlyingly structured, including many generalizations that do not hold for the language the child is trying to learn, or for any natural language for that matter. However, despite the fact that the data are insufficient to uniquely identify a generalization, children routinely arrive at the correct linguistic generalizations associated with their target language. The argument goes that this successful learning cannot be driven by patterns in the data — after all, the data are ambiguous — and must therefore be due to the child's innate learning preference. The perspective that children are guided by inherently endowed learning constraints is known as the **nativist perspective** on language acquisition (Clark & Lappin, 2010). Stimulus-poverty arguments also point to the rapidity with which children learn language as evidence that human infants do not entertain a large number of (eventually) incorrect hypotheses about their language. This suggests, again, that children are driven by inborn learning biases.

Neural networks, in particular neural language models, can inform this argument by offering one type of empirical evidence against stimulus-poverty claims (Lappin & Shieber, 2007). If an artificial learner can acquire the correct generalizations about a language without any linguistically informed learning biases, that suggests that, in principle, this is possible for a human language learner as well. Such evidence does not conclusively prove that children learn language without an innate learning bias. Rather, it serves as an existence proof and blueprint for what features of language could be learned via a domain-general, flexible learning model. For a deeper discussion of the role of neural network modeling in stimulus-poverty claims, see the discussions in McCoy, Frank, and Linzen (2018), Warstadt and Bowman (2020), Wilcox, Futrell, and Levy (2023a), and Yedetore, Linzen, Frank, and McCoy (2023).

How does the scaling paradigm impact the extent to which language models can inform stimulus-poverty arguments? We argue that neural language models can only *disprove* stimulus-poverty claims if they do not benefit from crucially relevant advantages that are unavailable to humans, with respect to both their inductive biases and their training data (Warstadt & Bowman, 2022). If the network has access to resources that are not available to human learners, then successful learning of a particular linguistic phenomenon no longer implies that this phenomenon is learnable, in principle, by humans. How humanlike in practice does a model learner have to be for its behavior to bear on stimulus-poverty claims? This is an active area of debate. There are some areas where language models have a clear disadvantage compared to humans; for example, during pretraining, they only observe language passively, and cannot benefit from interactions with other agents. However, LMs also have advantages over humans, e.g., while children need to learn language from a continuous speech stream, LMs are provided with textual input that is already segmented into linguistically meaningful units. Our view is that access to segmented textual input does not confer enough of an advantage to discredit stimulus-poverty claims, especially those relating to the acquisition of syntax and semantics (Warstadt & Bowman, 2022). However, one area that *might* confer a superhuman advantage is the amount of training data supplied to the model: if the LM is exposed to many thousands of times more sentences than humans, it may encounter plenty of evidence as to the correct generalization that is not available to human learners. In order to retain the validity of stimulus-poverty claims, then, it is necessary (at a minimum) to train models on no more examples of the linguistic phenomenon in question than is typical for human language learning.[4]

An additional reason why scaled-up language models bear less on stimulus-poverty arguments has to do with training data genre and quality. Mainstream large language models are trained on datasets of text scraped from the internet. Those tend to include text in dozens of different languages, as well as a substantial amount of code (Dubey et al., 2024, 17% of the Llama 3 training corpus). Worse, the content of the training corpus is often a proprietary trade secret, or else poorly documented or hard to search; indeed, it is not unreasonable to assume that training corpora include many linguistics and cognitive science textbooks or articles that discuss issues of learnability and give key examples of the critical generalization patterns. While logically speaking one can imagine scaling models up on clean monolingual English text that excludes linguistics articles, in practice it is difficult to construct clean corpora that are large enough to train mainstream LLMs, and the organizations that can afford to train those models in practice do not have an incentive to do so.

In summary, while models trained at smaller data scales can play an important role in assessing stimulus-poverty claims, mainstream large-scale LMs are limited insofar as how they bear on questions of language learning in people. Here, the issue with scaling is not directly the fact that model training requires a large amount of compute; rather, the limitation is caused by the corresponding scaling up of training data.

Testing probabilistic theories of language processing

The second type of contribution uses language models to empirically test theories of language processing that rely on probability distributions over words. In particular, language models have been important for developing and refining theories for the role of probabilistic prediction in language processing. As an example, we will discuss the impact that language models have made on the development of **surprisal theory** (Hale, 2001; Levy, 2008). Because scientists first started recording language processing behaviors, it has been widely observed that words that are less predictable in context are more difficult to process (Ehrlich & Rayner, 1981; Staub, 2015). Surprisal theory formalizes this observation by hypothesizing that the effort it takes to process a word is a (linear) function of its information content, or **surprisal**, the negative log probability of a word in its context. Previously, surprisal theory was tested primarily using non-neural-network based *n*-gram models (Smith & Levy, 2013), and sometimes using probabilistic context-free grammar (PCFG) language model (Hale, 2001). While such studies provided important early validation of the theory, those that used *n*-gram language models had several limitations, the most important being that the models used to estimate probabilities had a fixed window length, meaning that words outside of this fixed context were not factored into the estimate. While PCFG language models do not suffer from the fixed window length issue, they are limited in another way: they need to be trained on syntactically annotated data, of which we only have a small amount.

The advent of neural-network-based language models allowed researchers to compute more accurate probability estimates, enabling a more rigorous empirical assessment of surprisal theory. As a result, the relationship between word-level probabilities and human language processing behaviors has seen a surge of interest in the last five years: Using estimates from language models, some studies have validated the linear relationship between word-level surprisal and reading time (Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023), while others have challenged this original finding (Brothers & Kuperberg, 2021; Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023; Meister, Pimentel, Haller, Jäger, Cotterell et al., 2021). Other studies have investigated the surprisal–reading time relationship for cases where people read grammatically incorrect or implausible material, finding that reading times and surprisal values are poorly matched in these cases (Arehalli, Dillon, & Linzen, 2022; Huang et al., 2024; Van Schijndel & Linzen, 2021; Wilcox, Vani, & Levy, 2021). Recent work has gone beyond word-by-word reading times and used estimates from neural network models to argue that probability-based measures underlie decisions to skip words during reading (Pimentel, Meister, Wilcox, Levy, & Cotterell, 2023) or regress to a previous word (Wilcox, Pimentel, Meister, & Cotterell, 2024). Looking beyond linguistic processing, studies have used neural-network-based architectures to investigate the relationship between statistical co-occurrence and syntactic structure (Futrell, Qian, Gibson, Fedorenko, & Blank, 2019; Hoover, Du, Sordoni, & O'Donnell, 2021). The common theme between all these works is that each uses neural-network-based language models to estimate underlying word-level probability distributions, which can then be used to better empirically test theories of language processing.

How does the bigger and bigger trend of language modeling put this type of contribution in jeopardy? As language models grow in terms of architecture size and training data, their predictions appear to diverge more and more from those of people. For example, it has been shown that language models memorize large passages of text from their training data and will often repeat this text verbatim during generation tasks (Carlini, Ippolito, Jagielski, Lee, Tramèr et al., 2023), something that people do not do during natural language production (although they are certainly capable of such tasks, e.g., actors memorizing a script). This tendency towards long-form memorization as well as some similar types of biases, such as memorizing details only in certain contexts (Yehudai et al., 2024), suggests that, while better at language modeling, bigger models are worse for providing humanlike probability distributions that can be used to further psycholinguistic theories.

A recent line of work has clearly demonstrated the disadvantage of bigger models when it comes to modeling incremental reading times. To do so, Oh and Schuler (2023) and Shain et al. (2024) measured different models' predictive power: how well surprisal values estimated from those models predicted human reading times. Earlier work had suggested that as models' ability to predict upcoming words improved, their predictive power also increased (Goodkind & Bicknell, 2018; Wilcox, Gauthier, Hu, Qian, & Levy, 2020), albeit not for all languages (Kuribayashi, Oseki, Ito, Yoshida, Asahara et al., 2021). However, Oh and Schuler and Shain et al. found that for more recent models, the trend reverses. In other words, many of the models released in the past few years, which achieve state-of-the-art performance on a variety of natural language processing tasks, perform *worse* than their smaller-scale counterparts at predicting human reading times (Oh & Schuler, 2023; Shain et al., 2024). One possible explanation for this finding is that very large models are better than humans at predicting low-frequency words, thus predicting faster reading times for these items than is observed in the human data (Oh, Yue, & Schuler, 2024).

Unlike in Section 'Stimulus–poverty arguments', where the limitation arose due to scaling dataset size, LMs' misalignment with human behavior is likely due to the *combination* of large model size and large dataset scale. That being said, this work suggests that, in practice, models that both have fewer trainable parameters and are fit on smaller datasets are optimal for the types of studies described above.

[4] Human language learners are exposed to approximately 3 to 7 million words per year (Gilkerson et al., 2017; Hart & Risley, 1995). Therefore, by the time a child turns 12, an age by which they will have achieved grammatical competence that is adult-like in many respects, they will have experienced up to 100 million words. In comparison, mainstream language models are trained on multiple orders of magnitude more data.

The downsides of scaling for natural language processing

The scaling paradigm has downsides not only for psycholinguistics but also for language technologies. We survey some of these downsides in this section.

Dataset issues: Opacity and controllability. The scaling paradigm requires models to be trained on ever larger datasets. Very large datasets have several undesirable properties: First, they are **opaque**, meaning that their properties are not well understood. Although there have been recent calls for better dataset documentation (Gebru et al., 2021), most state-of-the-art LLMs are trained on datasets that are proprietary and are therefore fully opaque, or understood at only a very high level. Even for projects that do release some of all of their pretraining data (Biderman et al., 2023b; Groeneveld et al., 2024; Scao et al., 2022), the sheer size of the data can make it challenging to get an overview of. For example, the creators of The Pile (Gao et al., 2020), a large publicly available pretraining corpus, report that it is about 97% English, but say they cannot provide a reliable estimate of which other languages are represented in the dataset. Second, very large datasets are not **controllable**: it is hard to manipulate the contents of the dataset, for example, to remove harmful or toxic language, or to perform controlled studies of the impact of adding or removing pieces of the training data on a LM's behavior. Because current datasets are so large, it is both expensive and time-consuming to modify them, and because they are so opaque, it is not guaranteed that any given manipulation will successfully change all of its intended targets. Consequently, researchers cannot make good guarantees about the behavior of models trained on them.

Barrier to entry and homogeneity. Scaling produces a high barrier to entry for what is considered cutting-edge language modeling research. The training budgets for large-scale language modeling projects run into the tens or hundreds of millions of dollars, due to the required personnel, computer hardware, and energy costs (Sevilla, Heim, Ho, Besiroglu, Hobbhahn et al., 2022; Strubell, Ganesh, & McCallum, 2019). This has the potential to result in homogeneity of research directions, as those who can afford to participate in this research tend to be large technology corporations. Additionally, such high-cost research can produce a risk-averse research culture, even within well-funded organizations: if it costs significant amounts of money and compute to produce large language models, research groups will be more likely to focus only on methods that are highly likely to succeed. This increases the likelihood of scientific stagnation.

There are several proposed ways to broaden and democratize language model pretraining, several of which suggest distributing the overhead of training across many groups of researchers (Dean et al., 2012; McMahan, Moore, Ramage, Hampson, & Arcas, 2017). One benefit of scaled-down pretraining over these other training approaches is that, rather than splitting the training cost between multiple parties, scaled-down pretraining simply *lowers the cost altogether*. This means that new architectures and methods can be prototyped and tested quickly and cheaply and that large teams are not necessary. The scaled-down approach also aligns with the goals of the Green AI movement (Schwartz, Dodge, Smith, & Etzioni, 2020), which aims to reduce the carbon footprint of conducting AI and NLP research. Scaled-down pretraining is a beneficial paradigm alongside other proposals as a way to democratize and broaden participation in NLP and machine learning research.

Our recommendations for human-scale language modeling

Scientific progress that takes advantage of the synergies between psycholinguistics and NLP will require a dedicated focus on data-efficient and human-scale language modeling. Below, we outline several concrete proposals for how this can be accomplished.

1. **A curated set of cognitively inspired training datasets.** We recommend creating standard training datasets of a size commensurate with the amount of linguistic experience available to humans (Linzen, 2020). These datasets should ultimately include not only text, but also audio, transcriptions of audio, and multimodal data, such as aligned text–image and text–video data. The data domain should resemble the input to children and, in the ideal setting, would be recorded entirely from children's environments.[5] These datasets should be well-documented (Gebru et al., 2021) and should be made publicly available under a permissive license that allows academic, nonprofit, and private-sector research.
2. **A curated set of standard trained models for psycholinguistics research.** We recommend training and releasing open-weights models that are easily accessible and available in multiple languages. These models should be trained on publicly available datasets whose properties are well known, such as the cognitively inspired ones described in the previous paragraph. Scripts should be available to easily extract word-level probabilities from these models, enabling broad access in the psycholinguistics and linguistics communities, including to researchers who do not have the computational infrastructure to train new models.
3. **Incentives for data-efficient and small-scale language modeling research.** Incentive structures should be developed to encourage research that explores data-efficient pertaining. Such incentives could include workshops or shared tasks, such as the BabyLM Challenge discussed below, but also special issues of journals dedicated to human-scale pretraining (such as the issue in which this article is published).

We note that these recommendations overlap to some extent with ongoing efforts in the NLP and machine learning communities intended to improve the scientific and social benefits of LMs. This includes calls for better documentation of LM training corpora (Gebru et al., 2021; Lhoest et al., 2021; Ostendorff, Suarez, Lage, & Rehm, 2024) and discussion of the risks and benefits of open models, both in terms of the algorithm used to train the model and the final set of model weights (Biderman et al., 2023a; Bommasani et al., 2024).

Incentivizing human-scale language modeling: The BabyLM challenge

We next report on an effort we undertook to realize the recommendations outlined above—the BabyLM Challenge shared task. A **shared task** is similar to a competition, except that in addition to specifying win conditions, organizers often provide additional resources that help participants and lower the barrier to entry. Furthermore, the goal of a shared task is not just to *win*, but also to produce insights that will benefit a broader research community. Shared tasks have been used successfully in the past to bridge linguistics, NLP, and machine learning. For example, previous shared tasks have asked entrants to use NLP technologies to predict eye gaze data (Hollenstein, Chersoni, Jacobs, Oseki, Prévot et al., 2021), and morphological inflection schema (Cotterell, Kirov, Sylak-Glassman, Yarowsky, Eisner et al., 2016). The BabyLM Challenge was held in December 2023 as part of the CoNLL conference (the SIGNLL Conference on Computational Natural Language Learning).

The objective of the BabyLM Challenge was to train a language model using the same number of words of English available to a typically developing child in the United States—under 100 million words.

[5] Egocentric audio–video recordings of children's environments are available (Sullivan, Mei, Perfors, Wojcik, & Frank, 2021), but the input to a child is several orders of magnitude larger than what has been collected so far.

The structure of the challenge was designed to produce several concrete benefits for the NLP and computational linguistics research community. To keep participants on a level playing field, we collected a dataset of 100 million words, meeting recommendation (1) above. To enter, participants trained and shared models, following recommendation (2). By hosting the challenge at a conference and awarding prizes for the best models, we followed recommendation (3). While the challenge was partially motivated by the psycholinguistics considerations discussed in Section 'The downsides of scaling for psycholinguistics', we also wanted it to be useful for NLP practitioners interested in training efficient, small-scale language models (perhaps for the reasons mentioned in Section 'The downsides of scaling for natural language processing'). Some of the features of the challenge were therefore chosen to balance this consideration with our cognitive modeling goals, as well as to provide a level playing field for participants and enable fair comparisons across submissions. Below, we describe the structure of the challenge, summarize the results, and discuss their implications for psycholinguistics research. We also briefly discuss two follow-up experiments that we ran to answer several outstanding questions raised by the challenge about specific model architectures; for a fuller discussion of these experiments, see Appendix A. For this article, we keep the discussion relatively high-level; we encourage readers to refer to the original call for papers (Warstadt, Choshen, Mueller, Williams, Wilcox et al., 2023a) or the challenge proceedings (Warstadt et al., 2023) for more information about the technical aspects of the challenge. Each of the systems submitted to the challenge was accompanied by a paper describing the system; all of these papers are available in the challenge proceedings.

Overall structure: the three tracks

Submissions to BabyLM were required to conform to one of three guidelines, termed **tracks**. The three tracks 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 for all software used in the pipeline. *Loose* track submissions were encouraged to train on data beyond the linguistic text data provided through the shared task, for example, by conducting additional training on speech audio signal, code, music, or visual input. Language model training can involve making several passes over its dataset, where each pass is called an **epoch**. For the challenge, participants were allowed to train for as many epochs as they wished: multiple passes were not counted towards the 100M or 10M budget. Whether performing multiple epochs in training is cognitively plausible is an open question. Humans do not, of course, receive repeated exposure to the same stimuli. But there is evidence that we repeat some of the information we process to ourselves, for example, in memory replay (Carr, Jadhav, & Frank, 2011). That being said, because the winning submission performed hundreds of passes over the training corpus, we performed an experiment investigating the impact of epochs and found that most of the benefits of repeated exposure to the data occurred in the first 20 epochs. We describe these experiments in Appendix A, and otherwise put aside the issue of the cognitive plausibility of multiple-epoch training.

Training corpus

A major contribution of the BabyLM Challenge was the training dataset, which we refer to as the BabyLM Corpus. Ideally, of course, our data would exactly reproduce the input received by a child. Because such datasets are currently not available, our goal in this project was to make a step in the direction of this ambitious goal. One compromise we made, for example, is that our corpus consisted only of written texts or transcriptions of spoken language, while children's language exposure comes primarily from auditory or visual input (the latter in the case of signed languages). We reasoned that a conventional textual training corpus, despite this limitation, would attract a larger number of participants to the challenge.

Language model training corpora typically consist of text downloaded from web pages, online resource sites such as Wikipedia, and forums such as Reddit. In addition, they often include a large amount of non-linguistic content, such as computer code (e.g., Dubey et al., 2024). The BabyLM Corpus deviated from this typical composition in several respects: First, the majority ($\approx 56\%$) of the pretraining corpus was sourced from transcribed or scripted speech. This choice was made because much of the input to the typical child comes from face-to-face interaction, either through speech or sign. Transcribed speech may be particularly relevant when it comes to 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).

Another consideration was the genre of the transcribed speech. 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 (Huebner, Sulem, Cynthia, & Roth, 2021; Pannitto & Herbelot, 2020; Perfors, Tenenbaum, & Regier, 2011; Reali & Christiansen, 2005; Yedetore et al., 2023). While there is wide variability across cultures in the quantity of child-directed speech that is available to children, as opposed to overheard adult-to-adult interactions (Cristia, Dupoux, Gurven, & Stieglitz, 2019), many researchers hypothesize that children will learn particular words or structures more quickly given access to simpler child-directed inputs (see, e.g., Foushee, Griffiths, & Srinivasan, 2016; Shneidman & Goldin-Meadow, 2012). That said, children are routinely exposed to adult-to-adult interactions, and the extent to which adults vary their language when speaking to children differs greatly between cultures and socio-economic groups (Cristia et al., 2019). Accounting for these considerations and the availability of high-quality child-directed speech/text, about 40% of the data in the BabyLM Corpus came from sources either intended for children or appropriate for children, including child-directed speech, children's books, educational videos, and simplified English. Due to the limited amount of data in these genres, the remaining 60% came from adult interactions or writing for adult audiences, including Wikipedia articles and selections of books from Project Gutenberg. For more detailed descriptions of the data sources and preprocessing, see Warstadt et al. (2023). For the *Strict-Small* training corpus, we kept the proportion of data sources the same, sampling 10% from each source.

Evaluation tasks

Alongside the corpus, we also provided a pipeline to automatically evaluate LMs on a wide range of linguistic tasks. The pipeline, which was released as a public code repository,[6] consisted of well-known NLP evaluation benchmarks. Our evaluation tasks came in two paradigms: The first — called **zero-shot evaluation** — relied on obtaining outputs from the pretrained models. In our case, all of our zero-shot evaluations came from the BLiMP benchmark (Warstadt et al., 2020), which consists of tasks that evaluate whether the language models' predictions are consistent with the syntactic structure of English. Tasks consist of several example sentences, each of which targets a particular phenomenon of English syntax, for example, subject–verb number agreement. 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. A model is correct on a given example if it assigns a higher probability to the correct sentence in the minimal pair (Marvin & Linzen, 2018). We also created a supplement to the BLiMP tasks, which tests phenomena not captured by BLiMP. Unlike the original BLiMP tasks, which were released ahead of time, this supplement was released two weeks before the submission deadline.

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

This held-out evaluation was intended to reward models that could generalize well to never-seen-before evaluations.

The second evaluation paradigm involved fine-tuning, where we adapt a pretrained language model to a specific task by continuing to train it on a small dataset. For example, a pretrained LM that was originally trained on word prediction may be fine-tuned to predict entailment relationships between sentences. This type of evaluation is useful because during fine-tuning one can change the training objective of the model, such that it can be adapted into a tool for assigning categories to an input or giving binary judgments. Our fine-tuning evaluations included a subset of the tasks included in GLUE and SuperGLUE (Wang et al., 2019; Wang, Singh, Michael, Hill, Levy et al., 2018), consisting of various NLP tasks. Most of these tasks involve fine-tuning the model to perform classification; given an input sentence, the model is expected to sort the input into one of two classes. An example of such a classification task is natural language inference (NLI), where a model is given a **premise** sentence and a **hypothesis** sentence and has to categorize the relationship between them as `entailment`, `contradiction`, or `neutral`. An example premise is *Three tall boys are playing soccer*, and a hypothesis is *Some boys play sports*. Other tasks used similar techniques to investigate related aspects of meaning.

An additional fine-tuning task we included was the Mixed Signals Generalization Set (MSGS; Warstadt, Zhang, Li, Liu, & Bowman, 2020b). For this task, models were fine-tuned on an ambiguous training set where the labels were consistent with both a linguistic generalization and a surface generalization. They were then evaluated on examples that disambiguate which generalization the model converged on (if any). Surface behavior meant models were generalizing based on things like sentence length, orthography, or whether or not the sentence contained a particular word; linguistic generalization included whether or not the sentence contained an irregular past-tense form, or whether it contained a control construction. MSGS evaluates models on the assumption that one would like models to be more sensitive to linguistic features than surface features, as a systematic preference for abstract linguistic properties would make them better learners of language.

To compute the aggregate score across tasks, we weighted BLiMP and the BLiMP-supplement together at 50% (weighting all sub-tasks equally), GLUE and SuperGLUE together at 30%, and MSGS at 20%. While we do not have a strong motivation for this particular weighting, we found that the identity of the winning system for each track was not very sensitive to the weighting.

Baseline and skyline models

We trained and evaluated three baselines transformer models: OPT-125M (Zhang et al., 2022), RoBERTa-base (Liu et al., 2019), and T5-base (Raffel et al., 2020). As a skyline reflecting the state of the art in 2023, we also used our pipeline to evaluate Llama 2 (Touvron et al., 2023) (the variant with 70 billion parameters), which is a larger model trained on a massive corpus. Due to computational constraints, we evaluated Llama on GLUE and SuperGLUE using in-context learning instead of fine-tuning.

Submitted systems and results

We received 31 papers and 162 models in total. Some participants submitted to multiple tracks; we show data for unique participants in Fig. 2. Results of all models are shown in Fig. 3.[7] The scores of the top-performing models in each track are detailed in Table 1. Note that MSGS scores and BLiMP Supplement scores are lower than BLiMP and GLUE scores. For the BLiMP Supplement, this is likely due to the nature of the supplement tasks, which target well-formedness at the discourse level, including turn-taking and question-answer congruence. The lower supplement scores suggest that these sorts of generalizations are difficult to learn in a data-limited setting. For MSGS, because this task has not been run with human subjects, it is harder to say what counts as a *low* or *high* score. What MSGS does show is that all of our high-performing models (except the *Strict-Small* McGill-BERT) have a preference for structural generalizations over surface-level generalizations and that this preference is similar to the preference exhibited by large, pretrained models (i.e., Llama 2 and RoBERTa-Base). We believe that these results are important because they give us information about how the BabyLMs are solving our other evaluation tasks. If models had gotten zero or negative scores on MSGS, this would suggest that their performance on BLiMP and GLUE was likely due to memorized surface-level generalizations. However, because they (nearly) all scored positively on MSGS, this suggests that their performance on these other tasks could be due to the truly linguistic generalizations they acquired during pretraining.

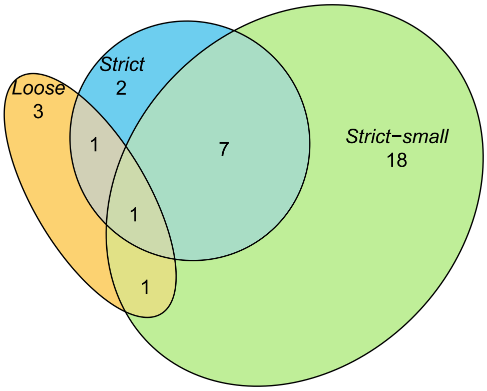


Fig. 2. Number of participants who submitted to each track, with multiple submissions counted once.

Below, in Section 'Common Methods', we break down the submissions based on the type of approach they use and discuss the effectiveness of these different approaches. Then, in the remainder of the section, we discuss the winning models in each track and what they can tell us about human language learning and processing. Before we discuss the details of any model or approach, we start by pointing out a few high-level takeaways from these results, beginning with comparisons between the different tracks. 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 performed better. 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 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 also close to achieving human-level performance on BLiMP and GLUE (i.e., they are near the green lines in Fig. 3). Interestingly, for BLiMP, the top-performing

[7] GLUE human scores are obtained by training crowd workers on each NLP task — for example, teaching them to classify entailment relations between sentences — as well as giving them 20 examples. For the BLiMP benchmark, human scores are obtained by asking naive participants to choose between sentences in a forced-choice task and calculating the proportion of times participants chose the grammatical variant.

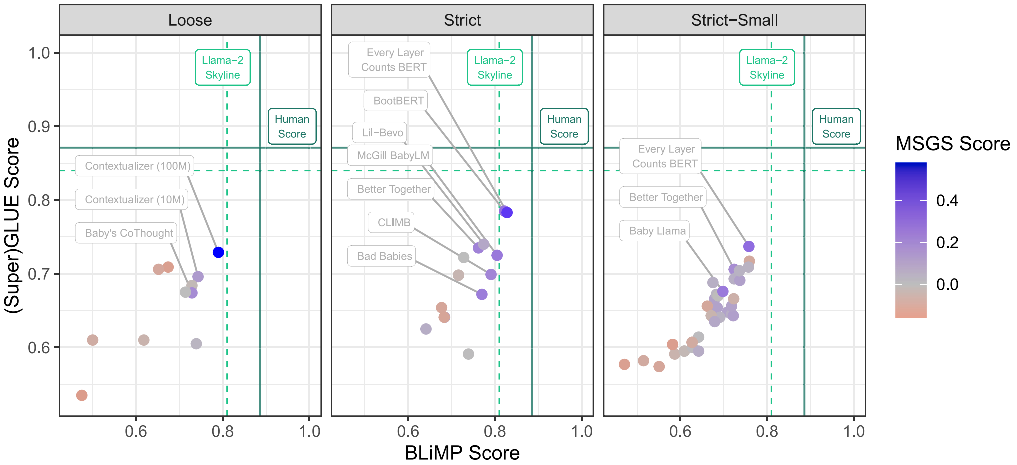


Fig. 3. Summary of BabyLM Submission Results: Each point represents an official model submission. Scores are broken down into performance on BLiMP (x-axis), GLUE and SuperGLUE (y-axis), and MSGS (color). Submissions that achieved an aggregate score above 0.6 are labeled in gray. Green dashed lines show Llama 2 skyline performance, and green solid lines show human performance. 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 reflects systematic linguistic generalization, and −1 is a systematic surface generalization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
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.

	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 & Samuel, 2023)	0.85	0.78	**0.47**	**0.77**	**0.74**
	BootBERT (Samuel, 2023)	**0.86**	**0.79**	0.28	0.72	0.70
	McGill-BERT (Cheng, Aralikatte, Porada, Piano, & Cheung, 2023)	0.84	0.72	0.25	0.71	0.67
	Best Baseline (OPT-125M)	0.75	0.70	0.13	0.68	0.60
Strict-Small	ELC-BERT (Charpentier & Samuel, 2023)	**0.80**	**0.74**	**0.29**	0.67	**0.66**
	MLSM (Berend, 2023)	0.79	0.71	0.17	0.57	0.61
	McGill-BERT (Cheng et al., 2023)	0.75	0.70	0.13	**0.68**	0.60
	Best Baseline (OPT-125M)	0.63	0.62	0.10	0.53	0.50
Loose	Contextualizer (Xiao, Hudson, & Al Moubayed, 2023)	**0.86**	**0.73**	**0.58**	0.63	**0.73**
	McGill-BERT (Cheng et al., 2023)	0.80	0.68	−0.02	0.57	0.57
	BabyStories (Zhao, Wang, Osborn, & Rios, 2023)	0.78	0.61	0.03	**0.65**	0.56

model is just a few percentage points shy of human performance. These results point to two important takeaways: (1) Human-level results have not been achieved *yet*. However, (2) connecting these results to our discussion in the previous section, we argue that the outcomes of the BabyLM Challenge bear on the stimulus-poverty arguments raised in Section 'Stimulus–poverty arguments'. While previous studies evaluating poverty of the stimulus (POS) claims have tended to use large, pretrained language models (e.g., Warstadt et al., 2020; Wilcox et al., 2023), the results of the challenge demonstrate that neural network learning algorithms are capable of learning linguistic generalizations, even when trained on human-scale datasets.

This being said, one challenge in connecting these results to POS claims is that our grammatical assay, BLiMP, tests many linguistic phenomena across several tasks, not all of which have been the locus of POS arguments. To allay this concern, we *post hoc* divided the subtasks based on whether they had been raised in debates on learnability in the previous literature. Our POS-relevant subtasks included ones that targeted *island constraints*, *filler–gap dependencies*, and *subject–aux inversion*. In Fig. 4, we compare average cross-submission performance on these tasks against all other subtasks in BLiMP. We find virtually no difference, suggesting that models are capable of acquiring generalizations about these hard-to-learn syntactic constructions.

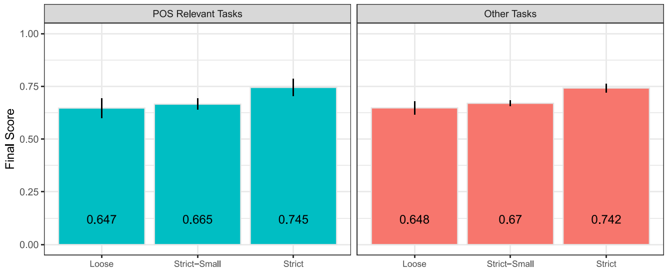


Fig. 4. BLiMP Subtask Performance: "POS Relevant Tasks", tasks that are relevant to poverty-of-the-stimulus debates, include ones that target filler–gap dependences, island effects, and subject–aux inversion. Error bars are 95% CIs across model scores on an individual task. Within each track, models on average perform similarly on POS-relevant and non POS–relevant tasks.

Common methods

To help us understand which approaches were effective, we hand-coded each submission based on the method(s) it employed. We show the breakdown of approaches in Fig. 5, and we visualize the trends

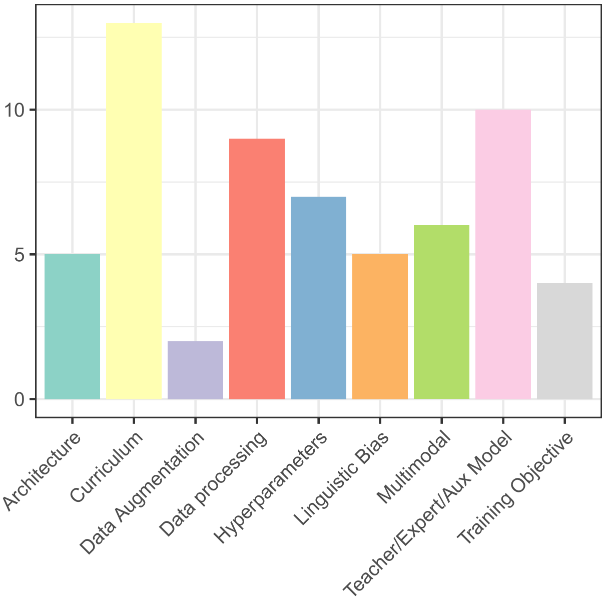


Fig. 5. Total number of submitted models that used each of the nine approaches in our typology. We count at most one submitted model per participant per track.

in Fig. 6. The three most important approaches for our purposes here were.

1. **Curriculum learning:** In curriculum learning, the corpus is sorted based on a metric — typically simplicity — and models are first trained on simpler examples before graduating to more difficult ones. The hypothesis is that this would lead to more effective or faster learning than presenting the data in a random order (Bengio, Louradour, Collobert, & Weston, 2009). Curriculum learning has some parallels to human language learning, in particular to child-directed speech, which is characterized by its reduced vocabulary size and simple constructions (Cameron-Faulkner, Lieven, & Tomasello, 2003). Some evidence suggests that child-directed speech helps language learning, especially with early vocabulary development and reading skills (Rowe, 2008); other work suggests that language learning proceeds at similar paces in groups where child-directed speech is not employed as frequently (Heath, 1983; Ochs, 1982). This approach can also be seen as related to the idea that successful learning depends on “starting small” (Elman, 1993).
2. **Data preprocessing**: Modifications to the underlying data, or the way the data is presented to the model with the exception of curriculum learning approaches.
3. **Architectural modifications**: This category includes systems that implemented changes to standard neural network architectures; we did not include methods that simply modified the default values of standard hyperparameters such as the learning rate.

While curriculum learning was the most popular approach in the submitted systems, it turned out to produce only marginal gains above the baselines. Data preprocessing and architectural modifications were found to be the most effective strategies in our meta-analysis.

All of the models submitted to the competition used a pre-existing **backbone architecture** (Fig. 7). All of the architectures were based on transformers (Vaswani et al., 2017), and many submissions were based on BERT (Devlin et al., 2019) and GPT (Radford et al., 2018). The backbone architectures differ in a number of ways, the most significant of these is that BERT and models derived from it, such as RoBERTa and DeBERTa, are masked LMs, meaning they predict a word given its surrounding context, whereas Llama and GPT are autoregressive LMs, meaning that they predict a word given only its preceding context. Overall, we find that models based on BERT, as well as several of its variants, including DeBERTa and LTG-BERT, achieved higher performance. In fact, the winning models for both the *Strict* and *Strict-Small* tracks used the LTG-BERT architecture. In the next sections, we discuss these winning submissions and ask what, if anything, they can tell us about human language learning or language processing.

ELC-BERT (method: architectural modification)

Architectural modifications. The winner of both the *Strict* and *Strict-Small* tracks was ELC-BERT (Charpentier & 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, Kutuzov, Øvrelid, and Velldal (2023). LTG-BERT combines four modifications to the Transformer architecture, all of which are relatively minor and were introduced in earlier papers: disentangled attention, enhanced layer normalization, GEGLU feed-forward modules, and scaled-down weight initialization. The most interpretable one is *disentangled attention*, drawn from DeBERTa (He, Liu, Gao, & Chen, 2021). The attention mechanism (Bahdanau, Cho, & Bengio, 2015), a central component of the transformer architecture, updates the representation of a word based on the representations of the other words in the context. In the original transformer, this mechanism has only indirect access to the position of the context words relative to the word whose representation is being updated. By contrast, disentangled attention explicitly factors the positions of each of the context words into the attention mechanism. The remaining three modifications are closer to the implementational level (in the sense of Marr’s (2010) levels of analysis), and it is therefore harder to assign an algorithmic-level interpretation to them (see Appendix B for details). ELC-BERT implements an additional modification on top of these four: whereas in a standard transformer the input to each layer is the output of the last one, in ELC-BERT the input to each layer is a weighted sum of the outputs of all previous layers (e.g., He, Zhang, Ren, & Sun, 2016).

Cognitive interpretation. What, if anything, can the success of the architectural modifications implemented in ELC-BERT tell us about human language learning? As mentioned above, most of the modifications concern implementation issues related to neural network optimization, and are difficult to interpret in cognitive terms; see Appendix B. That being said, one of the architectural modifications — disentangled representations of word position and word content — could plausibly introduce an inductive bias that makes it easier to learn abstract syntactic roles such as modifier or even subject, and as such could lead to stronger performance on benchmarks such as BLiMP; for a classic statement of the importance of separating roles and fillers in neural networks, see Smolensky (1990). Because ELC-BERT implements multiple simultaneous modifications on top of the standard transformer architecture, however, it is difficult to determine how much of its success can be attributed to this particular modification. In future work, this issue can be addressed with a controlled experiment that keeps the low-level modifications constant and varies only the type of attention mechanism used by the model.

Number of epochs. Apart from these architectural modifications, the submissions based on LTG-BERT stand out in that they were trained for many more epochs than other submissions. In particular, Charpentier and Samuel (2023) train models for over 450 epochs for their *Strict* submission, and over 2,000 epochs for their *Strict-Small* submission, which is much higher than is standard practice. This introduces a confound—did ELC-BERT perform well because of architectural modifications or because it was trained for far longer than any other model? To investigate this question, we conducted a follow-up experiment, presented in detail in Appendix A, where we trained ELC-BERT and LTG-BERT for only 20 epochs. The models’ scores dropped slightly

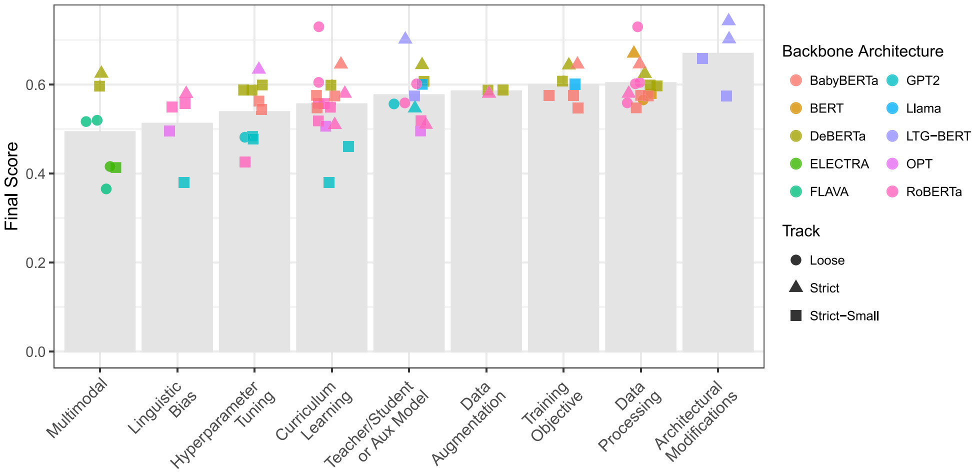


Fig. 6. Effect of Training Strategy and Backbone Architecture: Each point represents a submission. Some submissions may appear more than once if they use multiple strategies. Shapes show the challenge track to which the model was submitted. Colors show the backbone architecture on which the model is based. Gray bars show within-category aggregates. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

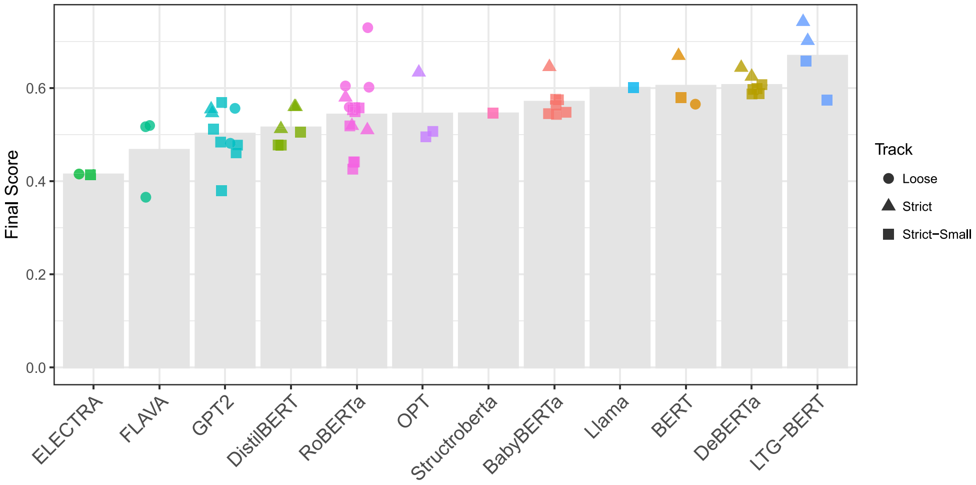


Fig. 7. Effect of Backbone Architecture: Each point represents a submission. Shape indicates the challenge track. Gray bars show within-category aggregates.

and were now in line with other top-scoring models. In other words, LTG-BERT can perform well with a much smaller compute budget than what was used in the team's submission, but when controlling for the compute budget, this architecture is not superior to others. We also found that LTG-BERT and ELC-BERT performed comparably, and concluded that it is the four modifications implemented in LTG-BERT architecture, rather than the skip connections — the additional modification introduced by ELC-BERT — that are responsible for the model's win.

Contextualizer (method: data augmentation)

The winner of the *Loose* track was the Contextualizer model of Xiao et al. (2023), which used a data augmentation scheme in which additional training samples are synthetically created by combining chunks of text from different sources in the dataset. Repeating this process 40 times for each chunk gives an augmented training corpus that has as many training samples as a four billion word corpus, but only uses 100 million words. Data augmentation is a common way to introduce inductive biases into models without changing their architecture. To make an image recognition system more robust to the precise location of the object in the image, for example, the system might be presented with many crops of the same training image; for a review, see, e.g., (Wang & Perez, 2017). In the case of language, in particular, repeating a syntactic constituent in many different environments is likely to help the learner recognize that the relevant sequence of words is a constituent; for example, if the learner observes the noun phrase *the big blue ball* as both a fragment answer to a question and also as a fronted element in a sentence, this provides evidence for its status as a constituent. This type of data augmentation has been used to endow models with a bias towards compositionally by teaching them that chunks can be recombined in different ways (Andreas, 2020). The empirical success of this method suggests that such an inductive bias is helpful for the acquisition of syntax. At the same time, large-scale data augmentation is arguably a less cognitively plausible method to impart this bias, compared to approaches that maintain human-like dataset sizes.

McGill-BERT

This submission from Cheng et al. (2023) was runner-up in the *Strict* and *Loose* tracks. The authors improve over the original BERT model by modifying two features: First, they shorten the context window, so the model only learns more local relationships between words. Second, they modify the way that training examples are presented to the model, splitting up examples into individual sentences, rather than in chunks that may contain multiple sentences. Rather than telling us something about psycholinguistic processes, the authors propose that this regime is particularly well-suited to the BabyLM training corpus, in particular its CHILDES portion. Because we include only child-directed utterances of CHILDES and remove any intervening child-produced utterances, each sentence does not necessarily follow from the previous one. Therefore, learning to predict these sentences separately, rather than as a single cohesive unit, may constitute an easier learning task.

CLIMB: A negative result for curriculum learning

In addition to track winners, we gave several awards to outstanding papers, one of which was the paper that describes the CLIMB system, focused on curriculum learning (Martinez et al., 2023). The authors experiment with three factors: the size of the vocabulary, the difficulty metric used to sort the corpus to construct the curriculum, as well as the model's objective function. Martinez et al. find that none of the curricula they tested yielded widespread improvements across the evaluation tasks, suggesting that curriculum learning, at least in its current form, may not be an effective method to construct sample-efficient language models.

We take these overall negative results for curriculum learning as fitting into an ongoing debate about the role of data limitations in language learning. This debate goes back to Elman (1993), who suggested that networks that were limited in certain respects in the initial phases of learning might prove to be better learners. Inspired by theories from cognitive science about how memory limitations in children might benefit rather than impede learning (Newport, 1988), Elman tested the impact of both memory and data limitations in language model training. He found that a simple recurrent network can learn the patterns of English embedded clauses, but only if trained initially on simple sentences that did not include embedded clauses, or on networks that were initially memory-constrained. This gave rise to the "starting small" hypothesis, namely that training models on initially simple examples and slowly graduating to more complex examples could lead to improvements in model performance (Bengio et al., 2009).[8] However, subsequent work testing this hypothesis yielded mixed results: running similar tests on more realistic datasets, Rohde and Plaut (1999) do not find evidence that starting small is beneficial to performance. Rather, they found that withholding complex examples at the beginning of training can hinder language learning in connectionist models. We take the negative results of Martinez et al. (2023), as well as other BabyLM submissions, as being in line with the conclusions of Rohde and Plaut (1999): simplifying the early stages of neural network training does not result in better learning outcomes, at least for small-scale datasets.

Looking forward: sample-efficient language models and psycholinguistics

The 2023 BabyLM Challenge led to several concrete outcomes aligned with our recommendations for more human-scale language modeling: it drew attention to the problem of data-efficient models and provided a venue for dozens of participants to share ideas and resources. Still, the challenge was limited in several important ways, especially as far as its implications for cognitive science. Below, we discuss some ways in which these limitations can be addressed in future iterations of the challenge.

Languages beyond English. The BabyLM challenge was conducted in only one language, English. Whatever mechanisms enable rapid language learning in the human mind, they do so regardless of the particular language being learned. Moving forward, it is therefore essential to test computational models on a variety of languages to ensure that the observed gains in performance are not specific to particular typological features.

Multimodal learning. Children learn not only from language but also from sensory contact with the world and from interaction with their caregivers and with each other. These input modalities may increase the learner's sample efficiency when measured in the number of input words (Zhuang, Fedorenko, & Andreas, 2024). We have made a step towards assessing the contribution of multimodal learning to sample efficiency by creating a vision and language track in the second iteration of the BabyLM Challenge (Choshen et al., 2024; Hu et al., 2024).

More challenging evaluation tasks. Our main evaluation tasks used either zero-shot minimal pair tests (BLiMP) or fine-tuning (GLUE) to probe models' linguistic abilities. These tasks are similar in many ways to the tests linguists and cognitive scientists use to probe this knowledge in people, but they also run the risk of overestimating model abilities. For example, looking at Fig. 3, one might take the performance of ELC-BERT and conclude that this model is roughly equivalent to the Llama-2 skyline. This raises the question: why would one ever bother to train a 70-billion parameter LLM on two trillion words of data when a much smaller model performs equally well? The answer is, of course, that while our BabyLM models are close to large-scale LLMs on our evaluations, large LMs remain far superior at more challenging evaluations, especially those that require generating text. When generating from our BabyLMs, for example, models often produce text that is filled with repetitions and is sometimes nonsensical. BabyLM models are also poor at following instructions or learning from examples in their input, something that larger LMs excel at. While their relatively strong performance on BLiMP and GLUE indicates that our small-scale models have learned interesting generalizations about grammar, this should not be taken to suggest that they are, in general, equivalent to large-scale LLMs. Finding evaluation tasks that better capture the limitations of BabyLMs compared to large-scale LMs is a necessary step for future iterations of the challenge.

Incentivizing cognitively motivated submissions. Another limitation of the systems submitted to the challenge has to do with the impact of the findings on psycholinguistics. While some of the findings have a cognitive interpretation — for example, the negative results for curriculum learning can be linked to ongoing debates in psycholinguistics about the importance of child-directed speech (Heath, 1983; Ochs, 1982; Rowe, 2008), and, in particular, support skepticism that child-directed speech is necessary for effective language learning — it is less clear how to interpret most of the positive results. Take the winning architecture ELC-BERT, for example. It is possible to draw loose parallels between the disentangled attention implemented by this architecture and cognitive theories that highlight the distinction between fillers and roles. However, most of the other features of ELC-BERT are not cognitively inspired in any meaningful way. This includes not only this system's low-level modifications to the transformer architecture Appendix B, but arguably also the transformer architecture itself, which keeps representations of all context words in memory, in sharp contrast with humans' limited working memory capacity (Armeni, Honey, & Linzen, 2022). The performance of other successful models is likewise only loosely connected to cognitive theories. McGill-BERT, for example, achieved high scores largely by changing model hyperparameters rather than by cognitively motivated modifications. One possible reading of this outcome is that theories from cognitive science have little to contribute toward effective small-scale language modeling, and *vice versa*. This conclusion is too pessimistic, in our view. The reason for the dominance of transformer variants may be that they enjoy an engineering infrastructure ecosystem that is convenient and optimized for efficient training. In future iterations, one could consider creating a separate track that focuses on novel, cognitively motivated architectures and training settings, where systems would not need to compete with heavily optimized transformers.

[8] Note there is an important difference between the starting small hypothesis, which is about *data* limitations, and the cognitive hypotheses which initially inspired Elman (1993), which are about *memory* limitations in children.

A wider range of psycholinguistic evaluations. Our evaluation tasks impose an additional limitation for the challenge's relevance to psycholinguistics. Psycholinguists have developed many different paradigms for collecting diverse types of human language processing data. However, we did not select tasks that represented the full breadth of such paradigms. For example, although we argue that LLMs can contribute to research on human sentence processing in Section 'The downsides of scaling for psycholinguistics', we did not ask how well our BabyLMs could explain incremental sentence processing data, even though such studies are well-established in the previous literature (Goodkind & Bicknell, 2018; Wilcox et al., 2023). We *did* provide an optional evaluation for the BabyLM challenge, assessing how closely models' word learning tracks that of a human child. This age of acquisition (AoA) task was taken from Portelance, Duan, Frank, and Lupyan (2023). In it, language models' surprisals were converted into a predicted AoA score by asking how much they help in predicting the age of acquisition over word frequency and concreteness ratings. Although we released code to run this evaluation, only seven teams evaluated on the AoA prediction task. Future iterations of the challenge should strengthen the connection to psycholinguistics by including evaluations that directly compare BabyLMs against a broad set of psycholinguistic data collected from people, especially from children.

Post-hoc analysis of successful systems. While this first round of the challenge was limited insofar as it did not produce any specific insights about the cognitive mechanisms for efficient learning, what it *did* do was identify a population of models whose architectures can serve as candidates for such mechanisms. It is possible that future work can derive cognitive insight from post-hoc analysis of the successful systems, with the goal of studying, conceptually and mathematically, *why* the modifications implemented by those systems facilitate sample-efficient learning. We are hopeful that once models are understood at a deeper mechanistic level, connections can be drawn with specific theories in the cognitive science of learning. Mechanistic interpretability methods, including circuit discovery (Conmy, Mavor-Parker, Lynch, Heimersheim, & Garriga-Alonso, 2023; Wang, Variengien, Conmy, Shlegeris, & Steinhardt, 2023) and causal feature analysis (Bricken et al., 2023; Huben, Cunningham, Smith, Ewart, & Sharkey, 2024; Marks, Rager, Michaud, Belinkov, Bau et al., 2024), have largely been applied to understand performance on NLP tasks, rather than to draw comparisons between language models and human language processing. However, there is a small but expanding literature that mechanistically investigates language models on phenomena of interest to linguists, including incremental sentence processing (Hanna & Mueller, 2024), property inference (Rodriguez, Mueller, & Misra, 2024) and quantifiers (Geiger, Lu, Icard, & Potts, 2021). Such analysis work should be incentivized in future iterations of the challenge.

General discussion

Cognitive modeling with neural networks has played an important role in psycholinguistics and many areas of cognitive science. As neural network approaches get more and more powerful, neural network modeling stands to produce many more insights in the decades ahead. At the same time, it is important to take stock and to ask how trends shaping the development of these models will impact their ability to help us answer scientific questions about the human mind. This paper has attempted to do just that. We have argued that, while beneficial for producing more powerful models, the current trend of using a standard model (transformers) and scaling up model size and training corpus has several potential downsides for psycholinguistics research. We recommend that linguists, cognitive scientists, and computer scientists work together to produce shared resources that are more human-scale, including human-scale pretraining corpora and models, as well as venues that support research dissemination in this area. In addition to the potential scientific impact of small-scale language modeling, we believe that focusing on such models has the potential to lower the barrier of entry for participation in language model research for engineering applications, allowing for a wider and more diverse set of interested scientists to contribute.

We reported on the BabyLM Challenge, one effort undertaken by the authors to actualize these recommendations. The most significant finding from the challenge itself is that, even at smaller data scales, current neural network architectures are very close to achieving human-level performance on many linguistic tasks. The best-performing models from the challenge showed sensitivity to syntactic constraints on par with models several orders of magnitude their size, and were just a few percentage points shy of human-level performance on this task. This is a significant achievement. Given the rate at which language modeling performance has improved recently, it is likely that computational models — even ones trained on human-scale datasets — will show sensitivities to some syntactic constraints that are on par with humans. The challenge produced several concrete outcomes, including (i) the BabyLM Corpus, (ii) a series of small-scale models, and (iii) several lessons for best practices in small-scale language modeling. Finally, the number of participants who contributed to the first iteration of this shared task demonstrates the broad interest in the topic. We are optimistic that, by thinking critically and carefully about the connections between machine learning and cognitive science, computational modeling researchers will continue to contribute to psycholinguistics in the decades ahead.

CRediT authorship contribution statement

Ethan Gotlieb Wilcox: Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Conceptualization. **Michael Y. Hu:** Writing – original draft, Visualization, Software, Methodology, Investigation, Data curation. **Aaron Mueller:** Writing – review & editing, Writing – original draft, Supervision, Software, Resources, Data curation, Conceptualization. **Alex Warstadt:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Resources, Methodology, Data curation, Conceptualization. **Leshem Choshen:** Writing – review & editing, Writing – original draft, Supervision, Conceptualization. **Chengxu Zhuang:** Writing – original draft, Supervision, Software, Resources, Investigation. **Adina Williams:** Writing – review & editing, Supervision. **Ryan Cotterell:** Writing – review & editing, Supervision. **Tal Linzen:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

E.G.W. was supported by an ETH Postdoctoral Fellowship. M.Y.H. was supported by an NSF Graduate Research Fellowship and NSF Award 1922658. A.M. was supported by a postdoctoral fellowship awarded by the Zuckerman STEM Leadership Program. A. Warstadt was supported by an ETH postdoctoral fellowship. This work was supported in part by the New York University IT High Performance Computing resources, services, and staff expertise.

Appendix A. experiments

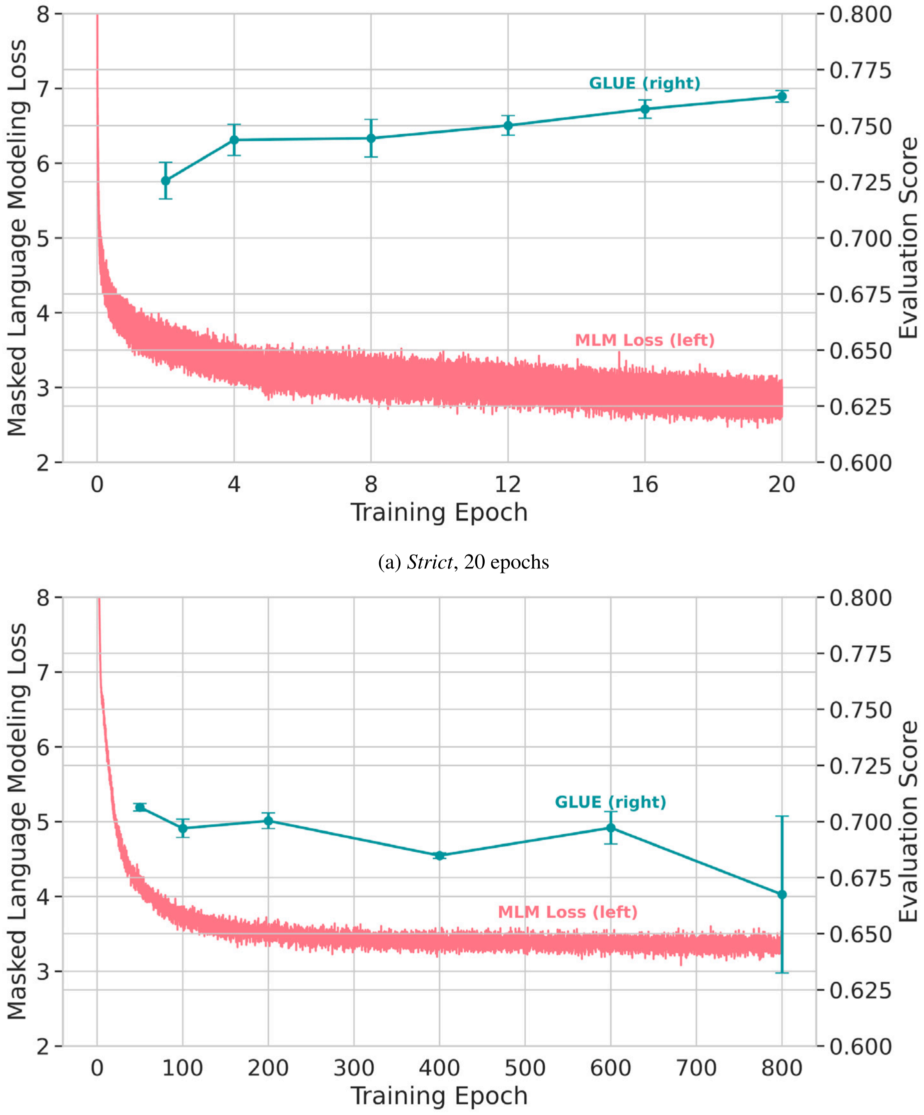
In this appendix, we present experiments inspired by unanswered questions from the BabyLM Challenge. First, we investigate the role of training time, measured in the number of epochs, on the performance of LTG-BERT. This architecture was originally trained with a significantly larger-than-standard number of epochs. Is such a large number of epochs necessary? Second, we directly compare ELC-BERT, which was the official winner of the BabyLM Challenge, against LTG-BERT, which is the backbone architecture on which it is based. We ask, are the skip connections between layers introduced by ELC-BERT necessary for strong performance in small-data language modeling? We find that LTG-BERT is about as good as ELC-BERT in our controlled setting, and that, while a large number of epochs can increase model performance, the returns diminish quickly with more epochs. We conclude that LTG-BERT is appropriate for successful small-scale language modeling and that it can be well-trained in about 20 epochs. However when trained for fewer epochs, it is not obviously better than other high-performing models in the challenge.

Data availability

Please see this repository, which contains code for training LTG-BERT and ELC-BERT models on the BabyLM training corpora.

Evaluating the role of the number of epochs in training

The BabyLM Challenge did not place any limits on the amount of computational resources participants could use when training their models. Because our dataset size was fixed for participants in the *Strict* and *Strict-Small* tracks, this meant that computational resources fluctuated as a function of (i) model size and (ii) training epochs, or the number of times the model sees its training data. Research in scaling has determined that training data size and model size should scale proportionally (Hoffmann et al., 2022); therefore, entrants tended not to train large models. When entrants did use more computational resources, this tended to be allocated toward an increased number of training epochs. When preparing baselines, we trained models for 20 epochs, which we chose based on prior experience. We intended



(a) *Strict*, 20 epochs

(b) *Strict-Small*, 800 epochs

Fig. A.8. Training curves and GLUE evaluation scores for *Strict* and *Strict-Small* LTG-BERT. All losses and scores are averaged over 3 random seeds. GLUE "Evaluation Score" is an average over all task-specific metrics (typically accuracy or F1-score). GLUE performance for *Strict-Small* LTG-BERT declines after training for 50 epochs. The Pearson correlations between training loss and GLUE performance are −0.97 and 0.61 for *Strict* and *Strict-Small* respectively.

Table A.2
A comparison between our reproductions of LTG-BERT and ELC-BERT (labeled "[R]"), our baselines, and existing results.

	Model	BLiMP	GLUE	MSGS	BLiMP-Supp.
	Llama 2	0.84	0.84	0.26	0.75
	RoBERTa-Base	0.87	0.79	0.24	0.76
Strict	ELC-BERT (Charpentier & Samuel, 2023)	0.85	**0.78**	**0.47**	**0.77**
	LTG-BERT (Samuel et al., 2023),	**0.86**	**0.78**	0.28	**0.77**
	[R] ELC-BERT, 20 epochs	0.83	0.75	0.25	0.67
	[R] LTG-BERT, 20 epochs	0.83	0.76	0.19	0.68
	Best non-LTG-based model (McGill-BERT)	0.84	0.72	0.25	0.71
	Best Baseline (OPT-125M)	0.75	0.70	0.13	0.68
S-Small	LTG-BERT (Samuel et al., 2023)	**0.80**	**0.74**	**0.29**	**0.67**
	[R] LTG-BERT, 800 epochs	0.76	0.67	0.02	0.63
	Best non-LTG-based model (MLSM)	0.79	0.71	0.17	0.57
	Best Baseline (OPT-125M)	0.63	0.62	0.10	0.53

this number—20 epochs—to also serve as a best first guess for our participants' training budgets, especially for those who did not have extensive prior experience training language models.

While most participants did indeed train in the general range of 20 epochs, some chose to train for much longer. In particular, the creators of ELC-BERT trained for 450 epochs in their *Strict* submission and 2,000 epochs in their *Strict-Small* submission, which is well beyond typical for language modeling research. Therefore, one big unanswered question at the end of the challenge was whether these models had achieved top scores because of their architectural innovations, or rather because they had trained for longer than other models. One other unanswered question from the challenge relates to the relative importance of the LTG-BERT baseline versus the skip connections introduced for ELC-BERT (described in Section 'Submitted systems and results'). Do the skip connections introduced in ELC-BERT significantly improve the model over and above the LTG-BERT baseline?

Methods

To answer these questions, we reproduced the LTG-BERT and ELC-BERT training pipeline using publicly available code from the authors and analyzed how the performance of the model improved over the course of training. We trained three models: For our first model, we trained LTG-BERT on the *Strict* dataset for 20 epochs to match our baselines. For our second model, we trained LTG-BERT on the *Strict-Small* dataset for 800 epochs, to more closely match the training epochs of the original LTG-BERT-based models submitted to the competition. Although we used fewer training epochs, our batch size was also smaller than the one reported in the original LTG-BERT paper due to computing constraints. Therefore, the number of gradient updates, or times when the model updates its weights based on the observed training data, is actually *higher* than that of LTG-BERT. For our third model, we trained ELC-BERT for 20 epochs on the *Strict* dataset. This was done so that we could make a direct comparison between ELC- and LTG-BERT when trained on the same number of epochs. Due to the significant cost of evaluating intermediate checkpoints, we only examine the final trained model for ELC-BERT.

The hyperparameters of the models trained in this experiment are given in Table A.3. All of the models trained for these reproduction experiments have a "[R]" next to their name. Note that our ELC-BERT reproduction uses the hyperparameters detailed in the *Strict* [R] column of this table. All training runs were done on 4 NVIDIA RTX8000 GPUs. The results of this experiment are shown in Table A.2.

Results

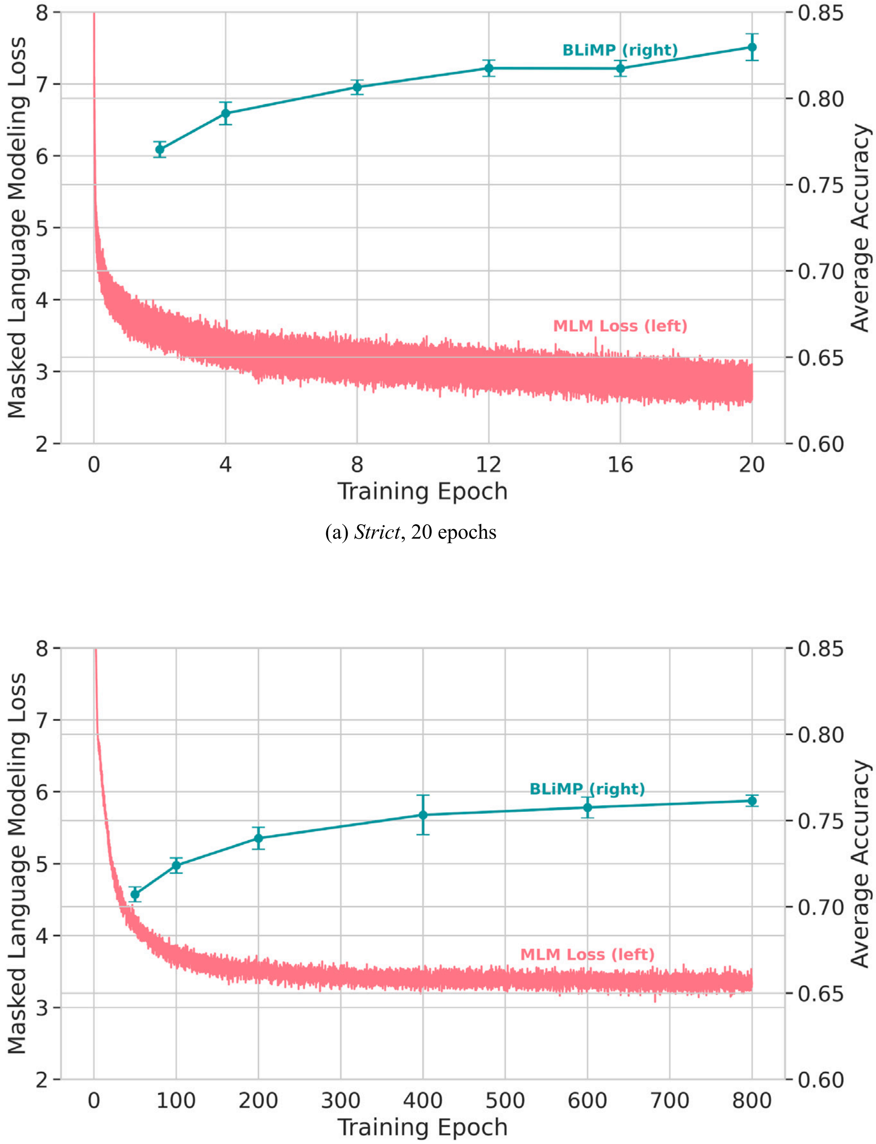
What is the impact of additional epochs? We find that both ELC- and LTG-BERT drop in performance when trained on only 20 epochs. Focusing first on our 100 million *Strict* models, the drop is about 2 percentage points on BLiMP and GLUE, about 10 percentage points on the BLiMP supplement, and an even larger decrease in correlation on MSGS. While we were not able to reproduce the training setup of ELC- and LTG-BERT exactly, these results suggest that a large number of epochs contribute to performance above and beyond model architecture. While still competitive, 20-epoch versions of ELC- and LTG-BERT are no longer clear winners of the *Strict* track. In A.2 we compare the scores of these models to the second-place model in the track, McGill-BERT (Cheng et al., 2023). While both 20-epoch models outperform McGill-BERT on GLUE, McGILL-BERT now gets better scores on BLiMP and the BLiMP supplement. For the *Strict-Small* track reproduction we observe similar results: Compared to the model trained on 2,000 epochs, the 800-epoch LTG-BERT performs 4 points lower on BLiMP and the BLiMP supplement, seven points lower on GLUE, and has a far lower correlation score on MSGS. Again, while there are still some differences in training details between us and the original authors, these results suggest that the number of epochs *does* contribute to the model's success. Compared against the second-place *Strict-Small* model, MLSM, the 800-epoch LTG-BERT model only performs better on the BLiMP supplement.

What is the impact of skip connections? Comparing our *Strict* ELC- and LTG-BERTs, we find that the two models are tied on BLiMP, ELC-BERT performs better on MSGS, but LTG-BERT has higher scores on GLUE and the BLiMP supplement (although only by 1 percentage point in each case). We interpret these results as indicating that the two models are approximately as good on our language-related evaluation tasks, and therefore that the skip connections of ELC-BERT do not add much above and beyond the original LTG-BERT architecture.

Training dynamics. In Fig. A.9, we visualize the training dynamics and BLiMP performance of models over training. We find that both the *Strict* and *Strict-Small* model's training loss decays roughly exponentially during training, as expected (Muennighoff et al., 2024). For both the *Strict* and *Strict-Small* models, the increase in BLiMP performance also diminishes exponentially over time. This trend also holds for the *Strict* model on GLUE (see Fig. A.8), but not for the *Strict-Small* model, where GLUE performance decreases slightly from 50 training epochs onwards. This pattern of diminishing returns on downstream tasks is in line with the previous literature on language model training (Hoffmann et al., 2022; Muennighoff et al., 2024). We do not observe any unusual learning dynamics, such as sudden drops in the training loss, nor instances in which test scores improve dramatically late in training, a phenomenon observed in some small-scale synthetic data experiments (Murty, Sharma, Andreas, & Manning, 2023; Power, Burda, Edwards, Babuschkin, & Misra, 2022).

Conclusions and recommendations

Our conclusions are threefold: First, it appears that in controlled settings ELC-BERT does not offer much of an advantage over LTG-BERT. We therefore recommend that practitioners interested in small-scale language modeling should use LTG-BERT, as this model is simpler. Second, when trained for fewer epochs, LTG-BERT is still an effective



(a) *Strict*, 20 epochs

(b) *Strict-Small*, 800 epochs

Fig. A.9. Training curves and BLiMP evaluation scores for *Strict* and *Strict-Small* LTG-BERT. All losses and scores are averaged over 3 random seeds. The Pearson correlations between training loss and BLiMP performance are −0.99 and −0.95 for *Strict* and *Strict-Small* respectively, indicating strong linear relationships. In other words, the training loss and BLiMP evaluations improve at roughly the same rate.

architecture for small-scale language modeling. Furthermore, because improvements decrease exponentially with additional epochs, we recommend that practitioners need not train this architecture for the number of epochs reported in Charpentier and Samuel (2023). Finally, it appears from our studies that while 20-epoch models are still effective, they are not clearly better than other top-performing BabyLM submissions. We, therefore, conclude that while LTG-BERT is a good architecture for small-scale language modeling, it was the large number of epochs that made it stand out above the other submissions.

Appendix B. LTG-BERT: Further details

Here we describe the three modifications to the standard transformer that are implemented by LTG-BERT:

1. Enhanced layer normalization. Standard layer normalization in transformers centers and scales the activations of a given layer to have zero mean and unit variance, such that the inputs to the next layer are of similar magnitude (Ba, Kiros, & Hinton, 2016; Vaswani et al., 2017). This has been found empirically to speed up and improve training. Following Shleifer, Weston, and Ott (2021), LTG-BERT applies this operation at more points along the models' computation than in the original transformer architecture.
2. GEGLU feed-forward modules (Shazeer, 2020). Each unit in a feed-forward layer computes a weighted average of its input, which is then passed through a nonlinear activation function. The standard transformer uses the ReLU activation function, whose output is zero for negative inputs and linear for positive ones. GEGLU is a more complex feed-forward layer. Here, the feed-forward layers learn two sets of weights and bias terms. The output of one set is passed through the Gaussian Error Linear Unit (GELU) activation function (which is similar to ReLU,

Table A.3
Pretraining hyperparameters. Differences between our training runs (labeled "[R]") and the original are bolded.

Hyperparameter	*Strict*	*Strict* [R]	*Strict-Small*	*Strict-Small* [R]
Number of parameters	98M	98M	24M	24M
Number of layers	12	12	12	12
Hidden size	768	768	384	384
FF intermediate size	2 048	2 048	1 024	1 024
Vocabulary size	16 384	16 384	6 144	6 144
Attention heads	12	12	6	6
Hidden dropout	0.1	0.1	0.1	0.1
Attention dropout	0.1	0.1	0.1	0.1
Training steps	15 625	**110 000**	31 250	**32 000**
Batch size	32 768	**256**	8 096	**2048**
Initial Sequence length	128	128	128	128
Final Sequence length	512	**128**	512	**128**
Warmup ratio	1.6%	1.6%	1.6%	1.6%
Initial learning rate	0.01	**3e-3**	0.005	0.005
Final learning rate	0.001	0.00141	0.005	0.005
Learning rate scheduler	cosine	cosine	cosine	cosine
Weight decay	0.1	0.1	0.4	0.4
Layer norm ϵ	1e−7	1e−7	1e−7	1e−7
Optimizer	LAMB	LAMB	LAMB	LAMB
LAMB ϵ	1e−6	1e−6	1e−6	1e−6
LAMB β_1	0.9	0.9	0.9	0.9
LAMB β_2	0.98	0.98	0.98	0.98
Gradient clipping	2.0	2.0	2.0	2.0

except it is curved around zero), and is then multiplied by the output of the other set.

3. The weights of the feedforward layers are initialized to values that are smaller than is standard, and that are progressively smaller for higher layers (Nguyen & Salazar, 2019).

Data availability

We have shared links to data and code in the article.

References

Andreas, J. (2020). Good-enough compositional data augmentation. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7556–7566). Association for Computational Linguistics, Online.

Arehalli, S., Dillon, B. W., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th conference on computational natural language learning* (pp. 301–313).

Armeni, K., Honey, C., & Linzen, T. (2022). Characterizing verbatim short-term memory in neural language models. In A. Fokkens, & V. Srikumar (Eds.), *Proceedings of the 26th conference on computational natural language learning* (pp. 405–424). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Ba, L. J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. CoRR, abs/1607.06450.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Y. Bengio, & Y. LeCun (Eds.), *3rd International conference on learning representations, ICLR 2015, San Diego, CA, USA, May (2015) 7-9, conference track proceedings*.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2024). Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, *121*(27), Article e2311878121.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning* (pp. 41–48). New York, NY, USA: Association for Computing Machinery.

Berend, G. (2023). 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).

Biber, D. (1991). *Variation across speech and writing*. Cambridge University Press.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., & Van Der Wal, O. (2023a). Pythia: A suite for analyzing large language models across training and scaling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of machine learning research*: *vol. 202*, *Proceedings of the 40th international conference on machine learning* (pp. 2397–2430). PMLR.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. (2023b). Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning* (pp. 2397–2430). PMLR.

Block, H.-D. (1962). The perceptron: A model for brain functioning. i. *Reviews of Modern Physics*, *34*(1), 123.

Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). Considerations for governing open foundation models. *Science*, *386*(6718), 151–153.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards monosemanticity: decomposing language models with dictionary learning. Transformer Circuits Thread. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of Memory and Language*, *116*, Article 104174.

Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*(2), 263–311.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems*: *vol. 33*, (pp. 1877–1901).

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction based analysis of child directed speech. *Cognitive Science*, *27*(6), 843–873.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). Quantifying memorization across neural language models. In *The eleventh international conference on learning representations*. openReview.net.

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

Charpentier, L. G. G., & Samuel, D. (2023). Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the babyLM challenge*. Association for Computational Linguistics (ACL).

Cheng, Z., Aralikatte, R., Porada, I., Piano, C. S.-D., & Cheung, J. C. K. (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).

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1979). The logical structure of linguistic theory. *Synthese*, *40*, 317–352.

Choshen, L., Cotterell, R., Hu, M. Y., Linzen, T., Mueller, A., Ross, C., Warstadt, A., Wilcox, E., Williams, A., & Zhuang, C. (2024). [call for papers] the 2nd BabyLM challenge: sample-efficient pretraining on a developmentally plausible corpus. arXiv preprint arXiv:2404.06214.

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157–205.

Clark, A., & Lappin, S. (2010). *Linguistic nativism and the poverty of the stimulus*. Hoboken, NJ: John Wiley & Sons.

Coffman, K. G., & Odlyzko, A. M. (2002). Internet growth: Is there a Moore's law' for data traffic? In *Handbook of massive data sets* (pp. 47–93).

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. In *Thirty-seventh conference on neural information processing systems*.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The SIGMORPHON 2016 shared Task—Morphological reinflection. In M. Elsner, & S. Kuebler (Eds.), *Proceedings of the 14th SIGMORPHON workshop on computational research in phonetics, phonology, and morphology* (pp. 10–22). Berlin, Germany: Association for Computational Linguistics.

Cristia, A., Dupoux, E., Gurven, M., & Stieglitz, J. (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, *90*(3), 759–773.

Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., Ranzato, Senior, A., Tucker, P., Yang, K., Le, Q., & Ng, A. (2012). Large scale distributed deep networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems*: *vol. 25*, Curran Associates, Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186).

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. (2024). The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.

Ehrlich, S. F., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 641–655.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*(1), 71–99.

Foushee, R., Griffiths, T., & Srinivasan, M. (2016). Lexical complexity of child-directed and overheard speech: implications for learning. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th annual meeting of the cognitive science society, CogSci 2016, Proceedings of the 38th annual meeting of the cognitive science society, CogSci 2016* (pp. 1697–1702). The Cognitive Science Society. Publisher Copyright: © 2016 Proceedings of the 38th Annual Meeting of the Cognitive Science Society, CogSci 2016. All rights reserved.; 38th Annual Meeting of the Cognitive Science Society: Recognizing and Representing Events, CogSci 2016 ; Conference date: 10-08-2016 Through 13-08-2016.

Futrell, R., Qian, P., Gibson, E., Fedorenko, E., & Blank, I. (2019). Syntactic dependencies correspond to word pairs with high mutual information. In *Proceedings of the fifth international conference on dependency linguistics (depling, syntaxfest 2019)* (pp. 3–13).

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., et al. (2020). The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92.

Geiger, A., Lu, H., Icard, T. F., & Potts, C. (2021). Causal abstractions of neural networks. In A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Kimbrough Oller, D., Hansen, J. H., & Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248–265.

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In A. Sayeed, C. Jacobs, T. Linzen, & M. van Schijndel (Eds.), *Proceedings of the 8th workshop on cognitive modeling and computational linguistics* (pp. 10–18). Salt Lake City, Utah: Association for Computational Linguistics.

Groeneveld, D., Beltagy, I., Walsh, E., Bhagia, A., Kinney, R., Tafjord, O., Jha, A., Ivison, H., Magnusson, I., Wang, Y., Arora, S., Atkinson, D., Authur, R., Chandu, K., Cohan, A., Dumas, J., Elazar, Y., Gu, Y., Hessel, J., ... Hajishirzi, H. (2024). OLMo: Accelerating the science of language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 15789–15809). Bangkok, Thailand: Association for Computational Linguistics.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the North American chapter of the association for computational linguistics*.

Hanna, M., & Mueller, A. (2024). Incremental sentence processing mechanisms in autoregressive transformer language models. CoRR.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of Young American children*. Paul H. Brookes Publishing.

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International conference on learning representations*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heath, S. B. (1983). *Ways with words: language, life and work in communities and classrooms*. Cambridge University Press.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., et al. (2020). Scaling laws for autoregressive generative modeling. arXiv preprint arXiv:2010.14701.

Hesslow, D., Zanichelli, N., Notin, P., Poli, I., & Marks, D. (2022). RITA: A study on scaling up generative protein sequence models. arXiv preprint arXiv:2205.05789.

Hestness, J., Narang, S., Ardalani, N., Diamos, G., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., & Zhou, Y. (2017). Deep learning scaling is predictable, empirically. arXiv preprint arXiv:1712.00409.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2022). An empirical analysis of compute-optimal large language model training. In A. H. Oh, A. Agarwal, D. Belgrave, & K. Cho (Eds.), *Advances in neural information processing systems*.

Hollenstein, N., Chersoni, E., Jacobs, C. L., Oseki, Y., Prévot, L., & Santus, E. (2021). CMCL 2021 shared task on eye-tracking prediction. In E. Chersoni, N. Hollenstein, C. Jacobs, Y. Oseki, L. Prévot, & E. Santus (Eds.), *Proceedings of the workshop on cognitive modeling and computational linguistics* (pp. 72–78). Association for Computational Linguistics, Online.

Hoover, J. L., Du, W., Sordoni, A., & O'Donnell, T. (2021). Linguistic dependencies and statistical dependence. In *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 2941–2963).

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, *7*, 350–391.

Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., & Wilcox, E. G. (2024). Findings of the second babylm challenge: sample-efficient pretraining on developmentally plausible corpora. arXiv preprint arXiv:2412.05149.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, Article 104510.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., & Sharkey, L. (2024). Sparse autoencoders find highly interpretable features in language models. In *The twelfth international conference on learning representations*.

Huebner, P. A., Sulem, E., Cynthia, F., & Roth, D. (2021). BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th conference on computational natural language learning* (pp. 624–646). Association for Computational Linguistics, Online.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. CoRR, abs/2001.08361.

Kuribayashi, T., Oseki, Y., Ito, T., Yoshida, R., Asahara, M., & Inui, K. (2021). Lower perplexity is not always human-like. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 5203–5217). Association for Computational Linguistics, Online.

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insightinto universal grammar. *Journal of Linguistics*, *43*(2), 393–427.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., ... Wolf, T. (2021). Datasets: A community library for natural language processing. In H. Adel, & S. Shi (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing: System demonstrations* (pp. 175–184). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

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

Linzen, T. (2020). How can we accelerate progress towards human-like linguistic generalization?. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5210–5217). Association for Computational Linguistics, Online.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*, 195–212.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., & Mueller, A. (2024). Sparse feature circuits: discovering and editing interpretable causal graphs in language models. CoRR, abs/2403.19647.

Marr, D. (2010). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press.

Martinez, R. D., McGovern, H., Goriely, Z., Davis, C., Caines, A., Buttery, P., & Beinborn, L. (2023). CLIMB – curriculum learning for infant-inspired model building. In *Proceedings of the babyLM challenge*. Association for Computational Linguistics (ACL).

Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 1192–1202). Brussels, Belgium: Association for Computational Linguistics.

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. In T. Rogers, M. Rau, J. Zhu, & C. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 2093–2098). Austin, TX.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In A. Singh, & J. Zhu (Eds.), *Proceedings of machine learning research*: *vol. 54*, *Proceedings of the 20th international conference on artificial intelligence and statistics* (pp. 1273–1282). PMLR.

Meister, C., Pimentel, T., Haller, P., Jäger, L., Cotterell, R., & Levy, R. (2021). Revisiting the uniform information density hypothesis. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 963–980). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., Pyysalo, S., Wolf, T., & Raffel, C. A. (2024). Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, *36*.

Murty, S., Sharma, P., Andreas, J., & Manning, C. (2023). Grokking of hierarchical structure in vanilla transformers. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 439–448). Toronto, Canada: Association for Computational Linguistics.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language. *Language Sciences*, *10*(1), 147–172.

Nguyen, T. Q., & Salazar, J. (2019). Transformers without tears: Improving the normalization of self-attention. In J. Niehues, R. Cattoni, S. Stüker, M. Negri, M. Turchi, T.-L. Ha, E. Salesky, R. Sanabria, L. Barrault, L. Specia, & M. Federico (Eds.), *Proceedings of the 16th international conference on spoken language translation*. Hong Kong: Association for Computational Linguistics.

Ochs, E. (1982). Talking to children in western samoa. *Language in Society*, *11*(1), 77–104.

Oh, B.-D., & Schuler, W. (2023). Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, *11*, 336–350.

Oh, B.-D., Yue, S., & Schuler, W. (2024). Frequency explains the inverse correlation of large language models' size, training data amount, and surprisal's fit to reading times. In Y. Graham, & M. Purver (Eds.), *Proceedings of the 18th conference of the european chapter of the association for computational linguistics (volume 1: Long papers)* (pp. 2644–2663). St. Julian's, Malta: Association for Computational Linguistics.

Ostendorff, M., Suarez, P. O., Lage, L. F., & Rehm, G. (2024). LLM-datasets: An open framework for pretraining datasets of large language models. In *First conference on language modeling*.

Pannitto, L., & Herbelot, A. (2020). Recurrent babbling: evaluating the acquisition of grammar from limited input data. In *Proceedings of the 24th conference on computational natural language learning* (pp. 165–176). Association for Computational Linguistics, Online.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338.

Pimentel, T., Meister, C., Wilcox, E. G., Levy, R., & Cotterell, R. (2023). On the effect of anticipation on reading times. *Transactions of the Association for Computational Linguistics*.

Portelance, E., Duan, Y., Frank, M. C., & Lupyan, G. (2023). Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal. *Cognitive Science*.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: generalization beyond overfitting on small algorithmic datasets. CoRR, abs/2201.02177.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67.

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

Rodriguez, J. D., Mueller, A., & Misra, K. (2024). Characterizing the role of similarity in the property inferences of language models. CoRR, abs/2410.22590.

Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

Rosenfeld, J. S., Rosenfeld, A., Belinkov, Y., & Shavit, N. (2019). A constructive prediction of the generalization error across scales. arXiv preprint arXiv:1909.12673.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, *35*(1), 185–205.

Rumelhart, D. E., McClelland, . J. L., Group, P. R., et al. (1986). *Parallel distributed processing, volume 1: explorations in the microstructure of cognition: foundations*. The MIT Press.

Samuel, D. (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).

Samuel, D., Kutuzov, A., Øvrelid, L., & Velldal, E. (2023). Trained on 100 million words and still in shape: BERT meets british national corpus. In A. Vlachos, & I. Augenstein (Eds.), *Findings of the association for computational linguistics* (pp. 1954–1974). Dubrovnik, Croatia: Association for Computational Linguistics.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., ... Thomas, W (2022). BLOOM: A 176b-parameter open-access multilingual language model. CoRR, abs/2211.05100.

Schaller, R. R. (1997). Moore's law: past, present and future. *IEEE Spectrum*, *34*(6), 52–59.

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, *63*(12), 54–63.

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). Compute trends across three eras of machine learning. In *2022 International joint conference on neural networks*. IEEE.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), Article e2307876121.

Shazeer, N. (2020). GLU variants improve transformer. CoRR, abs/2002.05202.

Shleifer, S., Weston, J., & Ott, M. (2021). Normformer: improved transformer pretraining with extra normalization. CoRR, abs/2110.09456.

Shneidman, L. A., & Goldin-Meadow, S. (2012). Language input and acquisition in a mayan village: How important is directed speech? *Developmental Science*, *15*(5), 659–673.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, *11*(1), 1–23.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1), 159–216.

Staub, A. (2015). The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass*, *9*(8), 311–327.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In A. Korhonen, D. Traum, & L. Màrquez (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3645–3650). Florence, Italy: Association for Computational Linguistics.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A Large, Longitudinal Audiovisual Dataset Recorded From the Infant's Perspective. *Open Mind*, *5*, 20–29.

Sutton, R. (2019). The bitter lesson. *Incomplete Ideas (Blog)*, *13*(1), 38.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). Llama 2: open foundation and fine-tuned chat models. CoRR, abs/2307.09288.

Van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), Article e12988.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December (2017) 4-9, Long Beach, CA, USA* (pp. 5998–6008).

Wang, J., & Perez, L. (2017). The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, *11*(2017), 1–8.

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*: *vol. 32*.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP workshop blackboxNLP: Analyzing and interpreting neural networks for NLP* (pp. 353–355). Brussels, Belgium: Association for Computational Linguistics.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2023). Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The eleventh international conference on learning representations*.

Warstadt, A., & Bowman, S. R. (2020). Can neural networks acquire a structural bias from raw linguistic data? In *Proceedings of the 42nd annual conference of the cognitive science society*.

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic structures in natural language* (pp. 17–60). CRC Press.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., & Zhuang, C. (2023). Call for papers - The BabyLM Challenge: sample-efficient pretraining on a developmentally plausible corpus. CoRR, abs/2301.11796.

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (2023). Findings of the babyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, & R. Cotterell (Eds.), *Proceedings of the babylm challenge at the 27th conference on computational natural language learning* (pp. 1–34). Singapore: Association for Computational Linguistics.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, *8*, 377–392.

Warstadt, A., Zhang, Y., Li, X., Liu, H., & Bowman, S. R. (2020). 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* (pp. 217–235). Association for Computational Linguistics, Online.

Whang, O. (2023). *The race to make A.I. smaller (and smarter)*. The New York Times, (Accessed 19 March 2024).

Wilcox, E. G., Futrell, R., & Levy, R. (2023a). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–44.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the cognitive science society*.

Wilcox, E. G., Pimentel, T., Meister, C., & Cotterell, R. (2024). An information-theoretic analysis of targeted regressions during reading. *Cognition*, *249*, Article 105765.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. (2023). Testing the predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*.

Wilcox, E., Vani, P., & Levy, R. (2021). A targeted assessment of incremental processing in neural language models and humans. 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)* (pp. 939–952).

Xiao, C., Hudson, G. T., & Al Moubayed, N. (2023). Towards more human-like language models based on contextualizer pretraining strategy. In *Proceedings of the babyLM challenge*. Association for Computational Linguistics (ACL).

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 9370–9393). Toronto, Canada: Association for Computational Linguistics.

Yehudai, A., Carmeli, B., Mass, Y., Arviv, O., Mills, N., Toledo, A., Shnarch, E., & Choshen, L. (2024). Genie: Achieving human parity in content-grounded datasets generation. arXiv preprint arXiv:2401.14367.

Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 12104–12113).

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M. T., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., & Zettlemoyer, L. (2022). OPT: open pre-trained transformer language models. CoRR, abs/2205.01068.

Zhao, X., Wang, T., Osborn, S., & Rios, A. (2023). BabyStories: Can reinforcement learning teach baby language models to write better stories? In *Proceedings of the babyLM challenge*. Association for Computational Linguistics (ACL).

Zhuang, C., Fedorenko, E., & Andreas, J. (2024). Visual grounding helps learn word meanings in low-data regimes. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: Long papers)* (pp. 1311–1329). Mexico City, Mexico: Association for Computational Linguistics.