Exploring smaller batch sizes for a high-performing BabyLM model architecture

Sharid Loáiciga, Eleni Fysikoudi, Asad Sayeed
Department of Philosophy, Linguistics and Theory of Science,
University of Gothenburg
sharid.loaiciga@gu.se, gusfysel@student.gu.se, asad.sayeed@gu.se

Abstract

We explore the conditions under which the highest-performing entry to the BabyLM task in 2023, Every Layer Counts BERT or ELC-BERT, is best-performing given more constrained resources than the original run, with a particular focus on batch size. ELC-BERT's relative success, as an instance of model engineering compared to more cognitively-motivated architectures, could be taken as evidence that the "lowest-hanging" fruit is to be found from non-linguistic machine learning approaches. We find that if we take away the advantage of training time from ELC-BERT, the advantage of the architecture mostly disappears, but some hyperparameter combinations nevertheless differentiate themselves in performance.

1 Introduction

The BabyLM Challenge (Warstadt et al., 2023a; Choshen et al., 2024; Charpentier et al., 2025) has become a shared task-style sandbox where researchers are invited to develop language models trained under developmentally plausible data budgets, simulating the linguistic input available to human children up to the age of 13. By setting small-scale amounts of data, either 10M or 100M words depending on the track, and providing standardized evaluation benchmarks, it aims to promote data-efficient modeling architectures. It also aims to support cognitively plausible approaches to automatic language acquisition. Finally, it is intended to broaden participation in language model research beyond large-scale industrial settings.

In this paper, we report on an expanded exploration of results for the ELC-BERT model (Charpentier and Samuel, 2023), the winning submission to the BabyLM Challenge 2023. This exploration focuses on the `strict-small` track, using the evaluation conditions and tools from the 2023 edition. Specifically, we investigate whether ELC-BERT's performance is primarily driven by computational resources, with a particular focus on batch size.

One of the findings of the 2023 edition of the shared task was that architectural innovations tended to be more successful than approaches inspired by curriculum learning or cognitive principles and the ELC-BERT architecture was at the forefront of that. ELC-BERT modifies the standard BERT architecture by replacing uniform residual connections with a learned weighting mechanism so that each layer selectively combines outputs from all preceding layers. This selective weighting means that the model can prioritize information from the most relevant layers, as opposed to treating all layers equally.

The approach achieved very strong performance, which makes ELC-BERT a great candidate as a base system going forward. However, it is also reported to have been trained for very long time and on a very large compute cluster. Our motivation in this work is primarily to investigate whether comparable performance can be achieved using substantially less computational resources using the ELC-BERT "tweak" to the BERT approach.

Since compute capacity is often a major limitation, both in terms of access and cost, efficient training methods are of particular importance. In this sense, it is worth noting that this year the BabyLM shared task has introduced stricter constraints on model training, limiting the number of epochs and training examples (Charpentier et al., 2025). This is a positive step toward standardizing experimental conditions and enabling more meaningful comparisons across replication studies.

Like all shared task submissions, ELC-BERT was likely developed under time constraints. We recognize that this is not an ideal setting for a comprehensive hyperparameter search. Our work addresses this gap by providing a more thorough investigation.

Our main contributions can be summarized as

Proceedings of the First BabyLM Workshop, pages 155–159

follows. In scenarios with more constrained computational resources:

- BLiMP scores are ultimately lower than the original ELC-BERT result.
- MSGS scores are higher than the original ELC-BERT result with a range of above-baseline outcomes on GLUE.
- A batch size of 32 with gradient accumulation of 12 (effective batch size of 384) achieves performance comparable to, or exceeding, that of larger batch sizes among the settings we tested.

All told, our experiments show that while ELC-BERT's original success depended originally in large part on the computational resources given to it, there are, nevertheless, some resource-constrained settings on which it does appear to be able to make performance gains.

2 Related work

Across both the 2023 and 2024 editions of the BabyLM Challenge (Warstadt et al., 2023b; Hu et al., 2024), the main insights remain consistent: language models can achieve good performance under strict data budgets, though substantial architectural innovations tend to yield greater gains than curriculum- or cognition-inspired methods. (In this case, "architectural innovations" are shorthand for alterations to Transformer-based machine learning approaches that are not directly inspired by linguistic or neurocognitive insights.) Efficiency in terms of data is the core objective of the task, but results also suggest a strong correlation between computational resources and performance, revealing a trade-off between data-efficiency goals and the benefits of larger-scale training.

Following these findings and with a focus on human-scale language modeling, Wilcox et al. (2025) examined this trade-off in high-performing systems from the shared task including ELC-BERT and LTG-BERT (Samuel et al., 2023), a closely related architecture from the same group that also performed very well. Importantly, they capped training at 20 epochs to control for the very long training of the original. The original ELC-BERT was trained for 31250 training steps and over 2000 epochs, whereas most other participants reported training for roughly 20 epochs. Wilcox et al. report performance comparable to the original, with only a 2-3 point drop in accuracy for both systems. However, specific results for the `strict-small` track for ELC-BERT are not provided (cf. Table A.2 in Wilcox et al.).

Furthermore, the ELC-BERT batch size for `strict-small` in both the LTG-BERT paper and ELC-BERT paper is 32768 with 128 tokens per sequence, totaling approximately 4M tokens. We estimate that this requires 2048 GB VRAM, equivalent to about 26 NVIDIA A100 GPUs. The batch size for the submitted ELC-BERT for `strict-small` is 8096, which we estimate requires about 6 A100 GPUs. In constrast, Wilcox et al. use a batch size of 2048 (cf. Table A.3) which requires 2 A100 GPUs. Lacking these computational resources, we investigate smaller batch sizes and use gradient accumulation as a means to approximate the bigger sizes. This of course also has an impact on the time that experiments run.

Within this context, we do not attempt a full replication of the experimental setup from the original ELC-BERT paper, as our computing infrastructure does not permit it. We focus instead on exploring the prospects for this type of architecture in more resource-constrained settings. These are arguably more plausible and faithful to BabyLM's attempt to simulate the conditions of human language acquisition *in silico*.

We performed a hyperparameter search within our computational constraints, contributing to transparency and supporting reproducibility. In this paper, we focus on batch size as representing one of the main resource bottlenecks of the ELC-BERT approach. We also contribute to the investigation of whether the original performance is obtained primarily from the "Every Layer Counts" architecture innovation or from the availability of substantial computational resources used to train ELC-BERT.

3 Set up

Most experiments in Table 2 were run on a single node of our computing cluster equipped with four NVIDIA Tesla A100 HGX GPUs (80 GB RAM each) and experiments with batches larger than 506 were run on 2 A100fat GPUs. Fine-tuning was performed on two GeForce RTX 3090 GPUs (24 GB RAM each). Pre-training used the hyperparameters specified in Charpentier and Samuel (2023), reproduced in Table 1. For fine-tuning, all hyperparameters from the official evaluation scripts[1] were

[1]https://github.com/babylm/evaluation-pipeline-2023

Hyperparameter	Submitted model
Number of parameters	24M
Number of layers	12
Hidden size is	384
FF intermediate size	1024
Vocabulary size	6 144
Attention heads	6
Hidden dropout	0.1
Attention dropout	0.1
Training steps	31250
Batch size	8096
Initial Sequence length	128
Warmup ratio	1.6%
Initial learning rate	0.005
Final learning rate	0.005
Learning rate scheduler	cosine
Weight decay	0.4
Layer norm ϵ	1e-7
Optimizer	LAMB
LAMB ϵ	1e-6
LAMB β_1	0.9
LAMB β_2	0.98
Gradient clipping	2.0
Gradient accumulation	1

Table 1: Pre-training hyperparameters for ELC-BERT model trained on the STRICT-SMALL track reported in Charpentier and Samuel (2023)

left unchanged[2].

4 Results and Discussion

Several observations can be made from Table 2. First, BLiMP scores, which focus on fine-grained grammatical knowledge, tend to be much lower in our replications compared to the original ELC-BERT results, even when training for the same number of steps. In contrast, although GLUE scores are also in general lower, the gap is not as wide. MSGS scores, however, always improve. Importantly, we note that GLUE and MSGS are obtained after a fine-tuning stage for which the default hyperparameters from the BabyLM evaluation set-up were used.

[2]In private communication with the original authors, we discovered that they were using an AMD-based architecture. On further investigation, we discovered that there are significant differences in the implementation of synchronization and gradient accumulation between AMD and NVIDIA that may have an effect on results.

A second observation concerns batch size and gradient accumulation. Runs using a batch size of 32 with gradient accumulation of 12 (effective batch size 384) achieves performance on GLUE and MSGS that matches or exceeds that of much larger batch sizes, while requiring significantly fewer computational resources. However, BLiMP performance seems insensitive to this and does not increase.

Training duration also emerges as an important factor. The original ELC-BERT was trained for over 2000 epochs, a scale of computation probably beyond what most academic teams can access. By comparison, our most efficient runs complete in minutes to a few hours, making them feasible for small groups or even individual researchers. This gap raises the question of how much infrastructure is required to remain competitive in modern NLP research and stresses the importance of computational budgets as well as data budgets.

Due to limitations in our computing infrastructure, we were unable to replicate the original batch size of 8096 and more than 2000 epochs. It remains an open question whether extended training can lead to convergence on different optima in the parameter space, given that language modeling does not converge toward a single optimal decision boundary.

5 Conclusions

We have evaluated the performance of ELC-BERT on a constrained setup and across several batch sizes, obtaining results that differ greatly from those reported in the original system description in a manner that suggests that the computational resources are, perhaps unsurprisingly, key to high performance even in data-constrained conditions. This is nevertheless significant because it *could* have been the case that the architectural innovation of ELC-BERT would have an much bigger influence on the outcome even under an environment of restricted computation.

In the BabyLM task, there is a healthy emphasis on comprehensive reporting of experimental conditions, including hyperparameters, training setup, and hardware specifications. Participants in the shared task complete a form where this information is reported. Future work in this area could involve incorporating such information into the benchmark itself, which could strengthen transparency and comparability across submissions.

Pre-training							Fine-tuning	
Batch size	**Training steps**	**Gradient accum.**	**Epochs**	**Time**	**BLiMP**	**BLiMP suppl.**	**GLUE**	**MSGS**
Original								
8096	31250	1	>2000	–	80.00	67.00	73.7	29.4
32	15625	1	4	21m	51.03	47.08	55.93	46.94
32	31250	1	7	44m	50.18	46.89	57.89	43.67
32	15625	12	41	2h39m	50.53	50.70	63.20	43.71
256	15625	1	27	1h7m	44.85	50.59	63.23	43.62
256	31250	1	53	19h57m	50.37	47.07	65.46	39.62
256	125000	1	218	8h31m	44.85	50.59	65.46	39.62
256	250000	1	437	17h4m	44.17	49.49	65.46	39.62
256	15625	12	333	5d10h5m	47.72	49.41	63.66	39.31
512	15625	1	55	1h49m	50.04	46.94	62.38	43.17
512	31250	1	109	3h37m	52.22	45.65	63.80	43.15
253	31250	32	1479	5d22h29m	46.95	49.88	63.72	39.31
506	31250	16	1736	3d18h42m	49.03	49.36	63.64	39.31

Table 2: ELC-BERT re-runs with varying batch sizes. Reported scores are average accuracies.

The shared task is still young, and beyond the work of Wilcox et al., there are very few analyses of this kind. We believe that such investigations are important for understanding both the reproducibility aspects and the broader implications of results in this setting.

A note on reproducibility The reproducibility crisis has been a subject of discussion for several years in addition to the usual pressures of the academic publishing cycle[3] (Baker, 2016), and computational linguistics and NLP are no exception. Despite its central role in scientific progress, reproducibility remains a persistent challenge in NLP research (Belz, 2022). In response, since 2020, reproducibility checklists are a requirement at submission time (Dodge et al., 2019; Magnusson et al., 2023), and initiatives such as ReproNLP, a shared task on reproducibility, have emerged (Belz et al., 2025). Nonetheless, the practical difficulties of reproducing scientific work are something that nearly every researcher eventually encounters.

In a shared task such as BabyLM, there is a risk that the whole task "overfits" on the most successful-seeming approaches on a year-to-year basis, in an environment in which the problem space is still not well defined (i.e., what even *is* an appropriate measure of human acquisition realism in language modeling?). Therefore, we argue that both replications *and* hyperparameter searches are core tasks in the BabyLM context, especially since the latter ensures that a result is placed in its proper theoretical context.

[3] https://www.nature.com/articles/d41586-024-04253-w

Limitations

Our computing infrastructure does not permit a direct replication of the original ELC-BERT training conditions. This limitation means that our work does not address reproducibility in the strict sense. Instead, it evaluates whether the original findings hold when training is conducted under scaled-down computational settings, and our conclusions should be interpreted within this narrower scope. Furthermore, no learning rate adjustments were made in conjunction with the smaller batch sizes, which may introduce bias into the results.

Acknowledgments

The work reported in this paper has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Addi-

tional funding was provided by the Gothenburg Research Initiative for Politically Emergent Systems (GRIPES) supported by the Marianne and Marcus Wallenberg Foundation grant 2019.0214.

The computations and data storage were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

The authors gratefully acknowledge the support provided by the developers of the ELC-BERT system and the organizers of BabyLM, particularly for their assistance with technical issues and specific inquiries.

References

Monya Baker. 2016. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Anya Belz. 2022. A Metrological Perspective on Reproducibility in NLP*. *Computational Linguistics*, 48(4):1125–1135.

Anya Belz, Craig Thomson, Javier González Corbelle, and Malo Ruelle. 2025. The 2025 ReproNLP shared task on reproducibility of evaluations in NLP: Overview and results. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 1002–1016, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, Raj Sanjay Shah, Alex Warstadt, Ethan Wilcox, and Adina Williams. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *Preprint*, arXiv:2502.10645.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts BERT. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 238–252, Singapore. Association for Computational Linguistics.

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candance Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. Call for papers – the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Michael Y. Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21, Miami, FL, USA. Association for Computational Linguistics.

Ian Magnusson, Noah A. Smith, and Jesse Dodge. 2023. Reproducibility in NLP: What have we learned from the checklist? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12789–12811, Toronto, Canada. Association for Computational Linguistics.

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

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

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Gotlieb Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Adina Williams, Bhargavi Paranjabe, Tal Linzen, and Ryan Cotterell. 2023b. Findings of the 2023 BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the 2023 BabyLM Challenge*. Association for Computational Linguistics (ACL).

Ethan Gotlieb Wilcox, Michael Y. Hu, Aaron Mueller, Alex Warstadt, Leshem Choshen, Chengxu Zhuang, Adina Williams, Ryan Cotterell, and Tal Linzen. 2025. Bigger is not always better: The importance of human-scale language modeling for psycholinguistics. *Journal of Memory and Language*, 144:104650.