BabyLLM: Comparison of BERT Masking Strategies Under Data and Capacity Limitations

Anonymous ACL submission

Abstract

Each year the LLM models expand in sizes which lead to the crisis of computational capacity and training resources. The BabyLLM challenge(Choshen et al., 2024) is designed to face these obstacles and question, whether we can still reach impressive results on model performance if we have to "bring up" it under the limitation of the average child development. The given paper reports the performance and experiment results under the open Strict Task for the 100M BabyLLM Challenge. We experimented with the choice of a Masking strategy that might improve the result and which strategies were better for which type of tasks.

1 Introduction

Pre-trained language models (PLMs) have emerged as a core paradigm in the field of Natural Language Processing (NLP), driving success across a wide range of tasks, with the BERT architecture being a key pioneer (Devlin et al., 2019). BERT models, which are deeply pretrained to learn bidirectional representations, capture rich contextual information while being fine-tuned for various tasks without substantial modifications to the architecture. A key component of BERT's pretraining technique is Masked Language Modeling (MLM) (Devlin et al., 2019). In MLM, some tokens in a given input sequence are randomly masked, and the model is trained to predict the original tokens based on the left and right context provided by the non-masked tokens (Yang et al., 2023).

While BERT's MLM approach has proven effective across various downstream NLP tasks (Devlin et al., 2019), it is computationally intensive and typically requires large-scale datasets to achieve state-of-the-art performance (Clark et al., 2020). However, in resource-constrained settings—where data availability and model capacity are limited—the effectiveness of different masking strategies becomes a crucial area of investigation. Despite the potential importance of optimizing masking techniques under these constraints, the exploration of how BERT's masking strategies perform in such low-resource environments remains under-explored.

1.1 Masking Strategies

Motivated by the BabyLM competition, this paper aims to address this gap by comparing various masking strategies. In addition to traditional BERT's MLM, we also include span masking and POS-Tagging Weighted (PTW) Masking.

Span Masking: Introduced in SpanBERT, this technique - in comparison with traditional masking - masks contiguous spans of tokens rather than individual tokens, improving the model's ability to understand longer phrases and syntactic structures (Joshi et al., 2020). By setting a novel span-boundary objective (SBO), the model learns to predict the entire masked span from the observed tokens at its boundaries.

POS-Tagging Weighted (PTW) Masking: This technique enhances model training by prioritizing difficult words based on their cumulative losses during training. It achieves this by weighting the masking probabilities of tokens according to their part-of-speech (POS) tags (Yang et al., 2023). By weighting the masking probabilities based on POS tags, PTW assigns higher weights to content words increasing their likelihood of being masked and enhancing the model's ability to focus on more informative and challenging tokens.

We evaluate the effectiveness of these strategies in low-resource environments, where both data and computational capacity are limited. Our objective is to observe how these different techniques can be leveraged to maintain strong performance in downstream tasks, even with fewer resources.

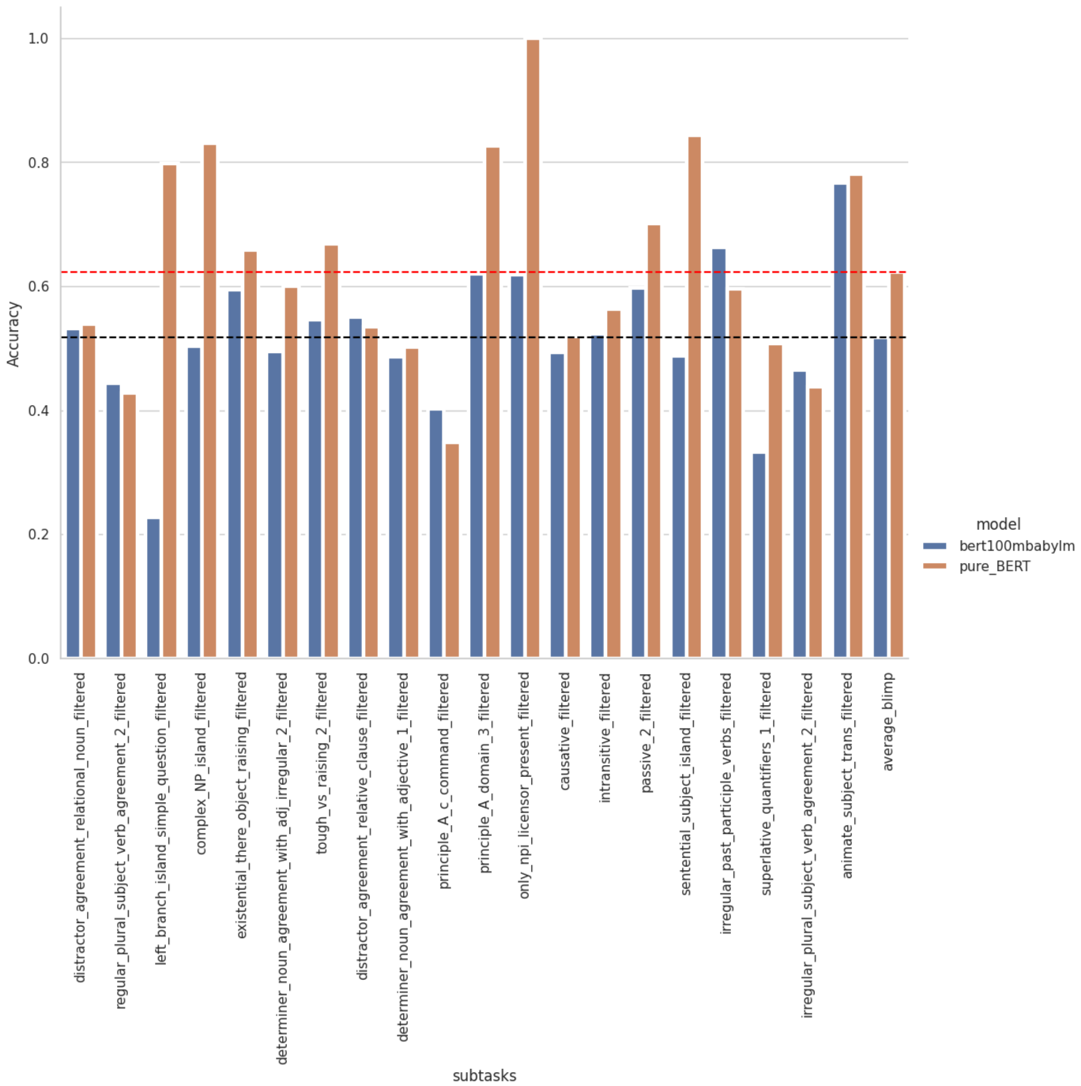

Figure 1: Blimp benchmark evaluation results in comparison between Standart Masking BERT and POS Masking Bert. On the x-axis, is a span of tasks out of Standard Blimp. A red line shows the average for the Standard BERT version performance, and a lack line for the POS BERT version respectively

2 Data

We trained our model with the 100M token dataset provided by the BabyLM competition, participating in the Strict track. The dataset includes six corpora: Open Subtitles, BNC Spoken, Gutenberg, CHILDES, Simple Wikipedia, and Switchboard. These corpora comprise a range of styles, from informal spoken language and child-directed speech to literary texts, offering a diverse training set for low-resource environments.

3 Evaluation

Our experiments were evaluated using a variety of benchmarks provided by the BabyLM Challenge:

BLiMP (Benchmark of Linguistic Minimal Pairs): This benchmark was used to assess the models' grammatical capabilities in the English language, through various language tests covering syntax, morphology, and semantics, such as anaphor-gender agreement and determiner-noun agreement. We also utilized the Supplement BLiMP, an extension of the original BLiMP benchmark, which includes additional and more challenging tasks not covered in the core BLiMP set (Warstadt et al., 2020).

GLUE (General Language Understanding Evaluation): A collection of natural language understanding tasks, GLUE evaluates models on a range of single-sentence tasks such as CoLA (linguistic acceptability) and SST-2 (sentiment analysis), as well as similarity and paraphrasing tasks like MRPC (Microsoft Research Paraphrase Corpus) and QQP (Quora Question Pairs), and natural language inference tasks including MNLI (Multi-Genre Natural Language Inference), QNLI (Question-Answering NLI), and RTE (Recognizing Textual Entailment) (Wang et al., 2018).

EWOK (Elements of World Knowledge): A benchmark designed to test the models' understanding of world knowledge across various domains, including social interactions and spatial relations. EWOK assesses how well a model can use its knowledge of concepts to distinguish between plausible and implausible contexts in a target text (Ivanova et al., 2024).

4 Results

We conducted an experiment that compared the evaluation results of the BERT model architecture under 3 masking strategies. Our main hypothesis is that guided POS masking will help to model performance. We trained BERT-large with the Standard Masking version and POS masking under the same hyperparams. Unfortunately, due to a lack of resources, we could not run LTG-BERT Span masking and use results released on Baby Challenge Github (Gao et al., 2023). Also, we should inform you that we evaluated BERT only on 10000 steps, acknowledging the fact mentioned on Github (Gao et al., 2023) that BERT architecture requires much more iteration for better performance in comparison to Decoder-only models.

4.1 Blimp

In Figure 1 We illustrate the chunk of Blimp benchmark performance of Standard BERT in comparison to POS Masking Bert. Based on the figure representations, we can conclude, that POS Masking and Standard Masking versions perform approximately at the same level most of the time. Meanwhile some tasks, like tasks on island or principle A, significantly outperform our POS Masking model. We could guess, that from the point of grammatical capabilities, POS tagging doesn't provide a significant impact on results. In Figure **??**, we also see the tendency that Standard BERT Masking outperforms POS Masking. Only on QA congruence easy task POS model outperforms Standard Masking. Results of LTG-BERT Span Masking were also included for correlation with the offered baseline.

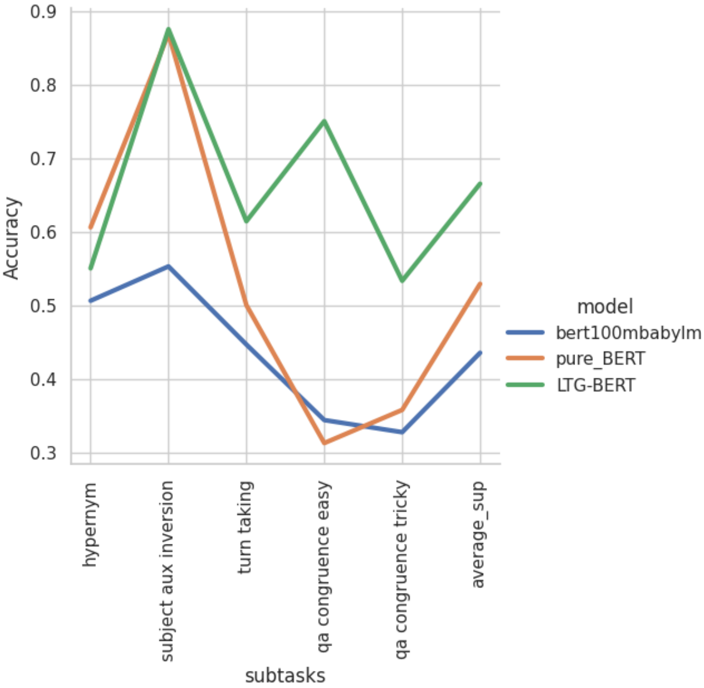

Figure 2: Blimp supplement benchmark evaluation results in comparison between Standard Masking BERT, POS Masking Bert and Span Masking LTG-BERT.

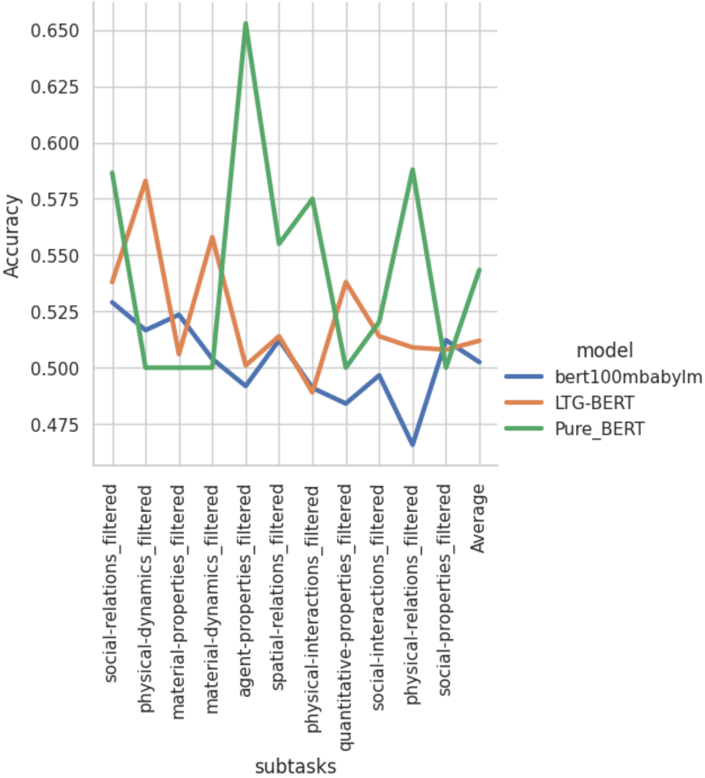

Figure 3: EWoK benchmark evaluation results in comparison between Standard Masking, POS Masking Bert and Span Masking LTG-BERT.

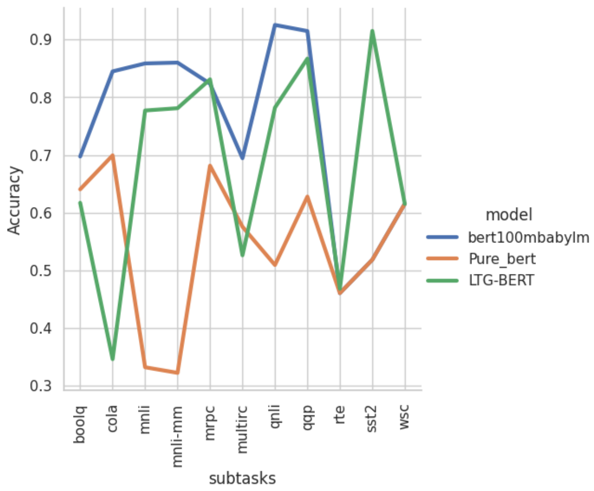

Figure 4: GLUE benchmark evaluation results in comparison between Standard Masking, POS Masking Bert and Span Masking LTG-BERT.

4.2 Ewok

In the case of the EWoK benchmark (See Figure 3), The POS solution performs better. Despite the lack of training, the model can reach or outperform the results of LTG-BERT span under 4 tasks. Surprisingly, but it seems that the Standard BERT solution reaches higher results than the more well-built LTG-BERT. This quite promising performance gives reasoning to state that longer training might help the model to build a better conclusion and explain its solutions more precisely.

4.3 GLUE

In Figure 4, the BERT with POS Masking shows impressive results on accuracy on 7 out of 11 tasks. The LTG-BERT reach 90 percent accuracy in the case of the SST 2 task, compared to 50 percent of POS Masking. The prominent accuracy is given under the COLA task, where LTG-BERT model accuracy is placed under 40 percent, meanwhile, the POS model is above 80 percent. The Standard approach performs lower in comparison to two other models.

5 Conclusion

Resolving the open strict task at BabyLM 2024 competition, we applied different masking strategies on standard BERT model, in order to find the effective training technique in low-resource environment. In the final submission we obtained highest number accuracy in GLUE benchmark for boolq, cola, mnli ,mnli-mm, multirc, qnli, and qqp subtasks in comparison with previous year winners LTG-BERT model. Given that our POS-tagging masking model was trained with limited number of training steps, it outperforms other models in tasks that requires fine-tuning. It thus proves that POS-tagging can enhance the model's ability to focus on informative tokens, which improved the performance on downstream tasks, even in resource-constrained scenarios.

Also, we observed that in certain grammatical evaluation tasks, such as the BLiMP benchmark, POS-tagging masking strategy achieved limited performance, indicating that standard masking strategies may still be effective for syntactic tasks. Therefore in future work, it is suggested that extending the training to more iteration steps could maximize performance across both grammatical and semantic tasks.

6 Limitations

Our experiment has faced several limitations. Firstly, our research was conducted on a smaller dataset (train100M.zip) due to the constraints of the BabyLM competition, which may limit the generalizability of our findings. Additionally, we used models with reduced capacity to investigate low-resource scenarios, potentially affecting the capture of complex linguistic patterns. The effectiveness of the masking strategies may also vary across different NLP tasks, and our focus on a specific set may not reflect broader performance. Furthermore, we performed minimal hyperparameter tuning due to limited computational resources, which could have impacted optimization. Our evaluation was confined to metrics defined by the competition, which may not fully capture the nuances of each strategy's performance.

References

Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. *Computing Research Repository*, arXiv:2404.06214.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *Preprint*, arXiv:2003.10555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Anna A. Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H. Clark, Carina Kauf, Jennifer Hu, R. T. Pramod, Gabriel Grand, Vivian Paulun, Maria Ryskina, Ekin Akyürek, Ethan Wilcox, Nafisa Rashid, Leshem Choshen, Roger Levy, Evelina Fedorenko, Joshua Tenenbaum, and Jacob Andreas. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *Preprint*, arXiv:2405.09605.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Preprint*, arXiv:1907.10529.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. pages 353–355.

Alex Warstadt, Yian Zhang, Haau-Sing Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). *Preprint*, arXiv:2010.05358.

Dongjie Yang, Zhuosheng Zhang, and Hai Zhao. 2023. Learning better masking for better language model pre-training. *Preprint*, arXiv:2208.10806.