AntLM: Bridging Causal and Masked Language Models

Xinru Yu[1*], Bin Guo[1*], Shiwei Luo[3*], Jie Wang[1], Tao Ji[2†], Yuanbin Wu[1†]

[1] School of Computer Science and Technology, East China Normal University
[2] School of Computer Science, Fudan University
[3] School of Computer Science and Technology, Harbin Engineering University

{xryu@stu,binguo@stu,jiewang@stu,ybwu@cs}.ecnu.edu.cn, taoji@fudan.edu.cn, shiweiluomo@gmail.com

Abstract

Causal Language Modeling (CLM) and Masked Language Modeling (MLM) are two mainstream learning paradigms based on Transformer networks, specifically the Decoder-only and Encoder-only architectures. The strengths of each paradigm in downstream tasks have shown a mix of advantages and disadvantages. In the past BabyLM Challenge 2023, although the MLM paradigm achieved the best average performance, the CLM paradigm demonstrated significantly faster convergence rates. For the BabyLM Challenge 2024, we propose a novel language modeling paradigm named **AntLM**, which integrates both CLM and MLM to leverage the advantages of these two classic paradigms. We chose the strict-small track and conducted experiments on two foundation models: BabyLlama, representing CLM, and LTG-BERT, representing MLM. During the training process for specific foundation models, we alternate between applying CLM or MLM training objectives and causal or bidirectional attention masks. Experimental results show that combining the two pretraining objectives leverages their strengths, enhancing overall training performance. Under the same epochs, AntLM$_{BabyLlama}$ improves Macro-average by 1%, and AntLM$_{LTG-BERT}$ achieves a 2.2% increase over the baselines.

1 Introduction

Language Modeling (LM) is a core task in NLP and a key technology for natural language understanding and generation, supporting a wide range of applications including machine translation (Hendy et al., 2023), speech recognition (Prabhavalkar et al., 2023), sentiment analysis (Tan et al., 2023), and information extraction (Wei et al., 2023). Over the past decades, LM has seen significant development, evolving from simple models like n-grams (Suen, 1979) to more sophisticated models, such as recurrent neural networks (Elman, 1990), long short-term memory networks (Hochreiter, 1997), and more recently, Transformer-based large language models (LLMs) like GPT (Radford et al., 2019) and BERT (Devlin, 2018). LLMs have demonstrated human-like or even superhuman performance in language modeling.

However, the tremendous success of LLMs relies on learning from massive corpora, which is not as data-efficient and low-energy as human language learning. The BabyLM Challenge 2023 (Warstadt et al., 2023a) and 2024 (Choshen et al., 2024) is a shared task over two consecutive years. It aims to encourage the discovery of more effective methods for training models using limited data. Considering that a 13-year-old child has encountered fewer than 100 million words in their lifetime, the shared task has introduced the *strict-small track*[1]. These tracks confine pre-training data to 10 million and 100 million words. These datasets consist of child-accessible materials, such as books, conversations, and Wikipedia entries, to enhance the relevance of language model pre-training to human language learning processes. Compared to 2023, the 2024 competition removed the Children's Book Test (Hill et al., 2016) and QCRI Educational Domain Corpus datasets (Abdelali et al., 2014). The 2024 competition also reduced the proportion of OpenSubtitles (Lison and Tiedemann, 2016) dataset while increasing the proportions of CHILDES (MacWhinney, 2000) and Project Gutenberg (Gerlach and Font-Clos, 2020) datasets.

The current investigation of LMs primarily adopts two predominant modeling paradigms: Causal Language Models (CLMs), represented by GPT (Radford et al., 2019), and Masked Language Models (MLMs), represented by BERT (Devlin,

---

[*] Equal contribution.
[†] Corresponding authors.

[1] Due to limitations in computational resources, we have not yet explored the *strict track* and the *multimodal track*.

2018). CLMs employ next-token prediction as their training objective, which is predicting the next token given the preceding context, and they perform exceptionally well on generative tasks. On the other hand, the training objective of MLM is the random selection and masking of some tokens in the input text, following which the model is trained to predict the original unmasked tokens. Due to its global information modeling capabilities, this approach excels in tasks necessitating the capture of bidirectional contextual information, such as text classification. Considering these modeling paradigms' strengths, this paper raises an important question: Could the two modeling methodologies be seamlessly integrated?

Intuitively, performing the MLM task allows the model to learn bidirectional contextual encoding of text, while the CLM task enables the model to predict and generate text based on prior content. These two learning objectives are not in conflict and could potentially be integrated. Analogous to a child learning a new language via practicing both cloze exercises and writing assignments, the training mechanism for a model can similarly employ a multi-task strategy. Therefore, we consider enabling our model to learn both tasks concurrently. To achieve this, we adopt a unified model architecture and alternate the training objective between MLM and CLM tasks. This approach attempts to mimic the human learning process, hence helping the model acquire deeper knowledge from a limited amount of text data.

To examine the effect of integrating MLM and CLM pretraining tasks on model performance, we conducted experiments using LTG-BERT and BabyLlama[2] as base models, testing on the BabyLM2024 10M datasets. LTG-BERT, an Encoder-only model, and BabyLlama, a Decoder-only model, are notable architectures from the 2023 BabyLM Challenge .The results indicate that both LTG-BERT and BabyLlama showed improvements in macroaverage scores. These experiments confirm that the integration of these two pretraining objectives can positively impact model training.

2 Related Work

Causal Language Models have played a pivotal role in the development of NLP, particularly in tasks involving sequence generation. The foundational work by OpenAI on the Generative Pre-trained Transformer (GPT) (Radford, 2018) marked a significant breakthrough in the use of CLMs for a variety of NLP applications. GPT (Radford, 2018) models the probability of each token in a sequence based on all preceding tokens, enabling it to perform well on tasks like text completion, machine translation, and summarization. The subsequent release of GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) further illustrated the power of scaling CLMs. These models, with their increased parameter sizes and training data, have set new performance benchmarks in tasks like zero-shot and few-shot learning. The GPT family firmly established the dominance of autoregressive models in generative tasks. More recently, Meta introduced the LLaMA (Touvron et al., 2023) series, which demonstrated that highly capable CLMs could be trained efficiently on fewer parameters and less compute than earlier models like GPT-3. LLaMA, designed to be accessible for academic research, retains the autoregressive framework while achieving competitive performance across a range of NLP tasks.

Masked Language Model is a training approach used to develop deep bidirectional representations of context, often referred to as a cloze task (Taylor, 1953). Specifically, a special token [MASK] is employed to randomly mask a proportion of input tokens, and the model is trained to predict these masked tokens. This training task was first innovatively introduced in BERT (Devlin, 2018) and has been adopted in subsequent models like RoBERTa (Liu, 2019) and ALBERT (Lan, 2019). Research has also led to improvements in MLM tasks, such as in SpanBERT (Joshi et al., 2020), where the model is trained to predict spans of words instead of individual tokens, enhancing its ability to capture long-range dependencies.

Unified modeling refers to using a single model architecture to handle multiple training and evaluation tasks. In the T5 (Raffel et al., 2020) model, various downstream tasks were reformulated as text-to-text tasks, significantly enhancing the model's ability for multitask learning. Moreover, many related works (Sanh et al., 2019; Liu et al., 2020) have also applied unified modeling for multitask training and evaluation, making it a common approach to improve the generalization ability of models. UniLM (Dong et al., 2019), based on the BERT architecture, is one of the significant endeavors in unified modeling. By employing specific self-attention

---

[2]We only utilized the BabyLlama architecture and did not apply the knowledge distillation method here.

masks, UniLM controls the contextual information used during prediction. When predicting tokens, it not only trains like an autoencoding language model by leveraging the context of masked tokens but also performs left-to-right training like an autoregressive language model. Additionally, UniLM can function similarly to encoder-decoder architectures by encoding the first input text and then generating sequences from left to right. By switching the attention matrix, it seamlessly transitions between different training tasks and downstream application scenarios.

Existing methods have unified CLM and MLM networks regarding model architecture and parameter sharing. However, research on unifying their training objectives remains unexplored. This paper is the first to bridge the two classic training objectives.

3 Methods

3.1 Preliminaries

BabyLlama (Timiryasov and Tastet, 2023) was proved to be effective in BabyLM2023 and is included as one of the baselines officially provided by BabyLM2024. BabyLlama (Timiryasov and Tastet, 2023) employed knowledge distillation, transferring the knowledge from two teacher models — a GPT-2 model with 705 million parameters and a LLaMA model with 360 million parameters — into a compact BabyLlama "student" model with just 58 million parameters. Given that our own replication of the BabyLlama model through distillation did not achieve ideal results, we opted to use only the BabyLlama architecture with a parameter size of 97 million. The BabyLlama model employs the classic CLM paradigm (Radford, 2018), where given the first n tokens in a sequence, the model predicts the token at position $n + 1$. The next-token prediction (NTP) training objective is to minimize the negative log-likelihood loss of predicting the next token at each timestep. To achieve this, a causal mask is applied in the self-attention mechanism. This mask is represented as a lower triangular matrix, ensuring each token can only attend to its preceding tokens. Formally, for an input sequence of length T, $x_1, x_2, \ldots, x_T$, the corresponding attention mask M is a $T \times T$ lower triangular matrix, where M_{ij} indicates whether the token at position i should attend to the token at position j This masking strategy effectively prevents the model from accessing future information during training, thereby capturing the sequential order and dependencies within the data.
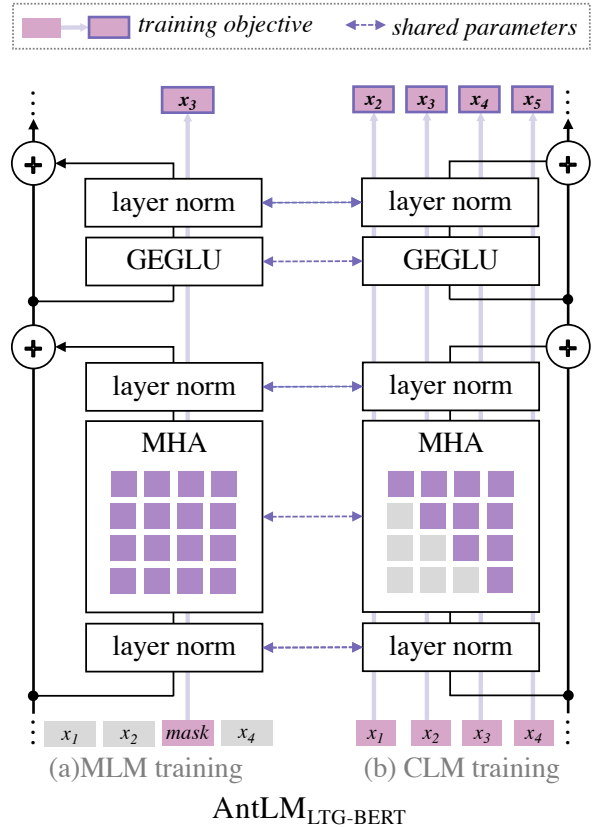

Figure 1: A diagram of AntLM$_{\text{LTG-BERT}}$. Based on the LTG-BERT architecture, we propose a joint MLM and CLM training objective. It is worth noting that the two objectives fully share parameters, but differ in their attention masks. The diagram also applies to AntLM$_{\text{BabyLlama}}$, with the difference in the architecture (e.g., positional encoding and the activation function of GLU).

In BabyLM2023 (Warstadt et al., 2023b), experiments with Boot-BERT (Samuel, 2023) and ELC-BERT (Charpentier and Samuel, 2023) demonstrated the effectiveness of the LTG-BERT (Samuel et al., 2023) architecture. LTG-BERT is also one of the official baselines in BabyLM2024. The LTG-BERT model incorporates several key architectural improvements, including NormFormer layer normalization (Shleifer et al., 2021), disentangled attention with relative position embeddings (He et al., 2020), and gated-linear activation function (Shazeer, 2020). The training objective of LTG-BERT is self-supervised Masked Language Modeling (MLM). During training, 15% of the tokens in the input sequence are randomly selected for replacement. Of these, 80% are masked, 10% are substituted with random tokens, and the remain-

ing 10% are unchanged. The model is then trained to predict the original masked tokens based on the context. LTG-BERT explores three common masking strategies: subwords, whole words, and spans. Experimental results indicate that span-based masking yields slightly better performance compared to the other methods.

3.2 Our Approach

Inspired by the way children learn languages through both cloze exercises and writing assignments, our work constructs a unified training framework that integrates CLM and MLM. In this unified framework, we switch between the two training paradigms alternately. CLM uses a causal mask to enforce sequential dependencies and MLM employs bidirectional attention, enabling the model to predict masked tokens by leveraging both preceding and succeeding context. By combining these two training objectives, the model not only excels at autoregressive tasks like text generation but also achieves a deeper semantic understanding of language by capturing broader contextual information through bidirectional attention.

In our approach, we integrate CLM and MLM by alternating between these training objectives during the pre-training phase. After training the model on one objective for a specified number of epochs, we switch to the other objective. The switch between training objectives is implemented by modifying the model's input and attention matrix. For the MLM task, 15% of the tokens in the input are randomly selected and replaced. The model utilizes bidirectional attention to predict the original tokens based on the surrounding context. In contrast, for the CLM task, no token replacement is required in the input. The model employs causal attention to predict the next token based on the preceding tokens.

4 Experiment

Data Preprocessing   For the data preprocessing part, we adopt the data handling procedures from the BootBERT (Samuel, 2023) method, which performed well in the previous round of BabyLM Challenge. Preprocessing includes steps like normalizing punctuation, reconstructing sentence structures, and removing duplicate text. These preprocessing steps help ensure cleaner and more structured input data, contributing to better model performance.

Name	BabyLlama	LTG-BERT
layers	12	12
attention heads	12	12
hidden size	768	768
intermediate size	2048	2048
vocabulary size	16k	16k
position bucket	–	32

Table 1: Model Hyper-parameters.

Baselines   We adopt the official baseline provided by the BabyLM Challenge as our benchmark, using the results achieved by the best-performing models from the previous round, namely LTG-BERT and BabyLlama, see Table 2.

Experiment Settings   In our experiments, we used both the BabyLlama and LGT-BERT models to evaluate the performance of a hybrid training strategy combining Causal Language Modeling (CLM) and Masked Language Modeling (MLM). For both model architectures, we used the same set of parameters, as shown in the table 1 and optimized the training process using the AdamW optimizer. Additionally, we employed the bfloat16 data type to enhance computational efficiency. For the BabyLlama model, we used a batch size of 512 with an initial learning rate set to 7×10^{-4}. The learning rate scheduler followed a cosine decay during the CLM training phase and a cosine with restarts scheduler during the MLM phase, with the number of cycles set to every four epochs . For the LGT-BERT model, we employed a batch size of 1024, with an initial learning rate of 5×10^{-4}. In all training phases, we used a cosine with restarts scheduler, with the num cycles set to 4. Our hyperparameters were determined through multiple experiments, building upon the hyperparameter settings from the previous works (Timiryasov and Tastet, 2023; Samuel et al., 2023) to find the optimal values. The training process alternated between CLM and MLM objectives over multiple epochs. We used the notation "x_CLM + y_MLM..." to indicate that, *in sequential order*, x epochs are trained in the CLM training mode, followed by y epochs in the MLM training mode, and so on.

4.1 Main Results

In this section, we evaluate the performance of BabyLlama and LTG-BERT across multiple bench-

Model	Data	BLiMP	BLiMP Supplement	EWoK	GLUE	*Macro average*
BabyLlama[†]	10M	**69.8**	59.5	50.7	63.3	60.8
BabyLlama	10M	68.1	60.4	50.4	65.5	61.1
AntLM$_{BabyLlama}$	10M	69.4	**60.7**	**51.1**	**67.4**	**62.1**
BabyLlama[†]	100M	73.1	60.6	**52.1**	**69.0**	63.7
LTG-BERT[†]	100M	69.2	**66.5**	51.9	68.4	64.0
BabyLlama	100M	**74.9**	66.0	52.0	66.3	**64.8**
LTG-BERT[†]	10M	60.6	60.8	48.9	60.3	57.5
LTG-BERT	10M	62.6	**65.4**	62.3	64.9	63.8
AntLM$_{LTG-BERT}$	10M	**72.3**	62.6	**63.0**	**66.0**	**66.0**

Table 2: Main experimental results. The † indicates results from the official report. The official BabyLlama leverages knowledge distillation, while our AntLM$_{BabyLlama}$ is based solely on the architecture of BabyLlama without knowledge distillation methods. Due to limitations in time and resources, we have not attempted AntLM on the 100M track, this will be part of our future work.

Training Stage	BLiMP	BLiMP Supplement	EWoK	Avg.
AntLM$_{BabyLlama}$				
8 (CLM)	68.2	56.7	50.5	58.5
16 (MLM)	56.8	58.4	57.2	57.5
24 (CLM)	68.1	60.4	50.4	59.6
24 (MLM)	56.9	57.8	**58.3**	57.7
4 (CLM) + 16 (MLM) + 4 (CLM)	**69.4**	**60.7**	51.1	**60.4**
AntLM$_{LTG-BERT}$				
12 (CLM)	69.9	56.4	50.8	59.0
60 (MLM)	62.8	**63.5**	64.2	63.5
72 (CLM)	70.0	57.2	51.9	57.9
72 (MLM)	69.4	61.1	**64.5**	65.0
6 (CLM) + 60 (MLM) + 6 (CLM)	**72.3**	62.5	63.0	**66.0**

Table 3: The effect of integrating CLM and MLM training objectives on BabyLlama and LTG-BERT.

marks, including BLiMP, BLiMP Supplement, EWoK, and GLUE. Our experiments primarily focus on assessing the impact of integrating CLM and MLM training objectives on the overall results, comparing the baseline performance of both BabyLlama and LTG-BERT with the configurations we propose.

As shown in Table 2, our models with integrated training objectives consistently outperform the official baseline scores on both the LTG-BERT and BabyLlama models. Notably, the improvements on LTG-BERT are particularly significant, demonstrating the effectiveness of our approach. To further validate the effectiveness of alternating training objectives CLM and MLM, we conducted an

in-depth experiment with the BabyLlama model. Given the lengthy training times associated with the GLUE dataset, we opted to evaluate our results on the BLiMP, BLiMP Supplement, and EWoK datasets. As shown in Table 3, the model trained with the *4_CLM+16_MLM+4_CLM* strategy significantly outperformed those trained solely with *8_CLM* or *16_MLM*. This finding indicates that combining these two training objectives enables the model to simultaneously acquire bidirectional context understanding and sequence generation capabilities. Under the same training epochs, the *4_CLM+16_CLM+4_CLM* combination demonstrated clear advantages over the pure *24_CLM* and *24_MLM* models, further confirming that the inte-
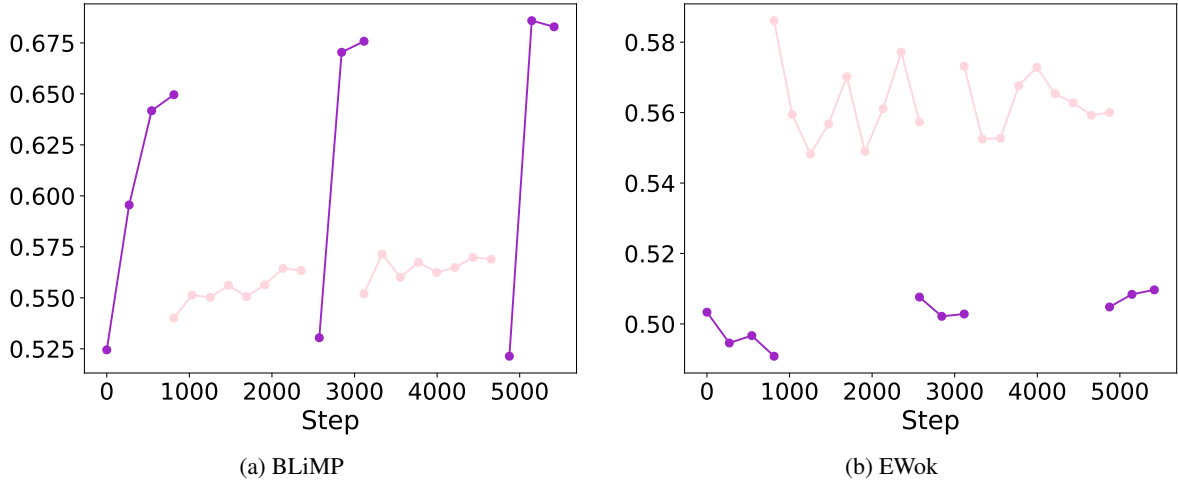
(a) BLiMP	(b) EWok

Figure 2: The phased experimental results on three datasets. The evaluation line chart for each stage of "3_ CLM + 8_ MLM + 2_ CLM + 8_ MLM + 3_ CLM " on the BabyLlama model. The reason for the discontinuity in evaluation results between training phases is that we applied the evaluation method corresponding to the specific task categories at each stage of the training process.

gration of these two training objectives is crucial for achieving optimal performance, highlighting the complementary relationship between CLM and MLM. We also conducted similar experiments on the LTG-BERT, the results are shown on same Table.

Additionally, we explored the performance of these training modes across different datasets. As shown in Figure 2, MLM performs significantly better on the EWoK dataset, while CLM exhibits more pronounced and sensitive results on the BLiMP dataset. This indicates that different training approaches have varying impacts on distinct datasets. Thus, the integrated experiments that combine both training methods can better leverage their strengths and enhance overall performance.

4.2 Ablation Study

To investigate the effects of various factors on the evaluation task results within the integrated experiments, we conducted ablation studies focusing on two variables: alternating frequency and alternating order. In the BabyLlama model, we maintained a constant total number of training epochs at 24 (8 epochs for the CLM phase and 16 epochs for the MLM phase). Specifically, for the alternating order, we adjusted the alternating sequence of training between the CLM and MLM phases while keeping the overall epoch count unchanged. For alternating frequency, we divided the training process into more frequent alternating stages. The experimental results, as shown in Table 4, indicate that varia-

tions in these two factors do not lead to significant declines in evaluation outcomes, suggesting that our approach is stable. We hypothesize that the decrease in performance with an increased frequency of alternations may be attributed to smaller epoch sizes in each training phase, which could hinder convergence on the respective tasks.

Furthermore, we found that the best performance was achieved when the CLM training phase was placed at both the beginning and the end of the training sequence, which could be due to the greater impact of CLM compared to MLM. Although CLM does not inherently have a higher performance ceiling (as last year's winner was an MLM-based model), but it converges more rapidly. CLM performs sequential prediction training on every token, while MLM focuses only on masked tokens. Thus, we suggest that CLM captures more learning within a single epoch than MLM.

5 Conclusion

In this study, we propose AntLM, a model that applies to multiple natural language-related tasks in the BabyLM Challenge by alternating between Causal Language Modeling (CLM) and Masked Language Modeling (MLM) during training. Experimental results demonstrate that AntLM achieves either superior or comparable performance to the baseline across all evaluation tasks.

Additionally, we found that CLM and MLM have different impacts on various evaluation tasks, suggesting that these training tasks guide the model
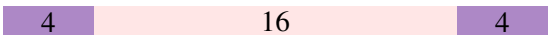
Training Stage	BLiMP	BLiMP Supplement	EWoK	Avg.
AntLM$_{\text{BabyLlama}}$				
8 · 16	68.2	56.7	50.5	58.5
16 · 8	68.4	**61.1**	50.1	59.9
4 · 16 · 4	**69.4**	60.7	**51.1**	**60.4**
8 · 8 · 8	67.2	59.2	50.2	58.9
4 · 8 · 4 · 8	68.8	60.6	50.7	60.0
8 · 4 · 8 · 4	68.6	59.1	51.0	59.6
3 · 8 · 2 · 8 · 3	69.3	60.1	50.8	60.1
1 3 1 3 1 2 1 2 1 2 1 2 1 2 1	67.3	55.2	50.4	57.6

Table 4: The effect of alternating frequency (low or high) and alternating order of CLM and MLM training objectives on BabyLlama. All were trained for a total of 24 epochs.

to learn distinct aspects of human language. We believe this difference is the key reason why integrated training yields effective results, as the model benefits from the knowledge learned from both training approaches. This finding also raises an intriguing question: do different training tasks allow models to capture only specific portions of natural language knowledge? Due to resource limitations, we were unable to explore additional ideas and approaches in this study. In future work, we plan to address these limitations by expanding our resources and support, allowing us to further investigate these potential directions.

Moreover, we conducted experiments with varying numbers and sequences of alternating training, and the results suggest that specific integrated training methods are more effective in achieving optimal evaluation outcomes.

References

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. 2014. The amara corpus: Building parallel language resources for the educational domain. In *LREC*, volume 14, pages 1044–1054.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Lucas Georges Gabriel Charpentier and David Samuel. 2023. Not all layers are equally as important: Every layer counts bert. *arXiv preprint arXiv:2311.02265*.

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. *Preprint*, arXiv:2404.06214.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

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

Martin Gerlach and Francesc Font-Clos. 2020. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1):126.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks principle: Reading children's books with explicit memory representations. *Preprint*, arXiv:1511.02301.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Z Lan. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8433–8440.

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

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. 2023. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Alec Radford. 2018. Improving language understanding by generative pre-training.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

David Samuel. 2023. Mean berts make erratic language teachers: the effectiveness of latent bootstrapping in low-resource settings. *arXiv preprint arXiv:2310.19420*.

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

Victor Sanh, Thomas Wolf, and Sebastian Ruder. 2019. A hierarchical multi-task approach for learning embeddings from semantic tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6949–6956.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

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

Ching Y. Suen. 1979. n-gram statistics for natural language understanding and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):164–172.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

Wilson L Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *Preprint*, arXiv:2308.02019.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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 Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors. 2023b. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Singapore.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.