ELC-ParserBERT: Low-Resource Language Modeling Utilizing a Parser Network With ELC-BERT

Rufus Behr
Research Computing, Northeastern University
r.behr@northeastern.edu

Abstract

This paper investigates the effect of including a parser network, which produces syntactic heights and distances to perform unsupervised parsing, in the Every Layer Counts BERT (ELC-BERT) architecture trained on 10M tokens for the 2024 BabyLM challenge. The parser network's inclusion in this setup shows little or no improvement over the ELC-BERT baseline for the BLiMP and GLUE evaluation, but, in particular domains of the EWoK evaluation framework, its inclusion shows promise for improvement and raises interesting questions about its effect on learning different concepts. [1]

1 Introduction

Recent advancements in Transformer-based language models, in particular Large Language Models (LLMs), have largely been achieved by scaling the parameter count as well as the size of the dataset (Zhao et al., 2023). Whilst there is ongoing research in identifying efficient training and sampling methods for LLM pre-training, Villalobos et al. project that between the year 2026 and 2032 the datasets for training LLMs will be equivalent to all extant human text data.

In response to the staggering amount of data upon which LLMs are trained, the BabyLM challenge aims to incentive research in the development and pre-training of Language Models by setting realistic human-developmental limitations on the training data (Choshen et al., 2024). In particular, the challenge has three data-limited tracks: two texts only tracks that restrict the data corpora sizes to 10M and 100M (strict-small and strict, respectively), the latter of which is inspired by approximately the amount of data a 13 year old child will have seen, and a vision-language track, combining text and images.

[1]The code for the training and experimenting is available here: https://github.com/SufurElite/ELC-ParserBERT

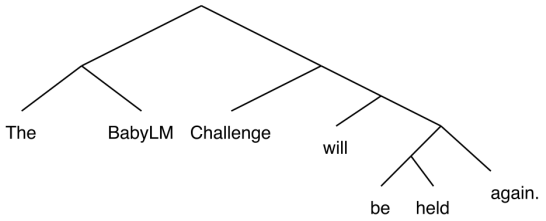

Figure 1: An example of an induced tree created from the model's unsupervised parser network

The 2024 BabyLM Challenge is the second iteration of this challenge. The overall best system from the first challenge was the Every Layer Counts BERT model (ELC-BERT) (Georges Gabriel Charpentier and Samuel, 2023), which showed effective results by changing the residual connection between the transformer layers. Although, in the first BabyLM challenge, systems with architectural modifications produced the best results, a plurality of submitted systems used curriculum learning, of which only one found significant gain from this approach (Warstadt et al., 2023).

This paper introduces ELC-ParserBERT, a model submitted for the strict-small track, which incorporates the parser network proposed in (Shen et al., 2021) into ELC-BERT. The parser network is able to induce both dependency and constituency syntactic structures, an example of which can be seen in Figure 1, and the aim of its inclusion is to investigate whether this structural bias aids the baseline ELC-BERT model. This paper also investigates whether using a curriculum learning based approach with this model architecture yields any improvement.

2 Background Literature

Hu et al. propose a Transformer-based Syntactic Language Model (SLM), called Generative Pretrained Structured Transformers (GPST), that learns to induce syntactic parse trees in an unsu-

The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, pages 140–146

pervised manner and is able to outperform GPT-2, including in the GLUE (Wang et al., 2018) evaluation dataset. In addition to the standard SLM with its Transformer backbone, the GPST has a composition component, a pruned inside-out encoder, namely, ReCAT (Hu et al., 2024b), which induces parse trees. The model is trained through a process akin to hard expectation-maximization: during the expectation stage, the model induces a parse tree from a compositional model, whose internal representation is used as input during the maximization stage that consists of updating all the GPST parameters.

The ReCAT component and its contextual inside-out layers made improvements upon unsupervised grammar induction when compared to the prior baselines (Hu et al., 2024b). One of the baselines it improved upon (both in terms of the F1 score for the syntactic trees and the memory complexity) was the StructFormer model (Shen et al., 2021).

The StructFormer also proposes an additional component, the parser network, that induces parse trees. Given its input of word sequences, the parser network generates syntactic heights and distances, which were proposed in (Luo et al., 2019) and (Shen et al., 2018), respectively. Given the syntactic heights and distances, the network then estimates the probability that a token is the head of another token. A directed weighted adjacency matrix is then created such that each weight is the probability a token depends on another. After these token-dependency relation probabilities are created in the parser network, they are used to constrain the self-attention (Shen et al., 2021). In addition to being evaluated on its unsupervised dependency and constituency parsing, the StructFormer was trained and evaluted as a masked language model.

As part of the first BabyLM Challenge, one system consisted of pre-training the original StructFormer architecture as well as variants of it (Momen et al., 2023). Their variants included using RoBERTa encoder in place of the standard transformer and at which layers to integrate the parser network – namely, placing the parser in the middle since there was supporting literature that shows syntactic information is better represented in the middle transformer layers. They concluded, however, that, although some of the evaluation tasks were improved upon by having a model that induces a syntactic bias into the architecture, there was not sufficient evidence that this inclusion improved the model architecture with respect to the challenge set nor, within their experimentation, that the placement of the parser network in the middle of the transformer layers yielded improvement.

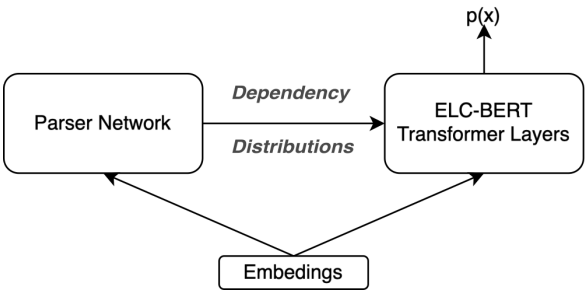

Figure 2: The model architecture

Another approach to inducing grammar induction is through compound probabilistic context-free grammars (compound-PCFGs) (Kim et al., 2019), wherein the model's context-free rule probabilities are alterable by a sentence-level latent variable. There was a submission to the first BabyLM challenge that made use of the compound-PCFG. The approach pre-trained a compound-PCFG on a subset of the strict-small training data and used the token embedding layer from the grammar as the input embedding layer to a different language model, which is subsequently trained on next work prediction on the training data (Chen and Portelance, 2023). They concluded, however, that there was no improvement over their baselines on account of the grammar induction, but that their choice of tokenizer, which was the WordPiece algorithm used to create both subword and whole word tokens, may have resulted in increased performance.

As mentioned above, the best submission to the first BabyLM challenge – and one of the baselines for this iteration – was the ELC-BERT (Georges Gabriel Charpentier and Samuel, 2023), which did not try to leverage syntactic structures but rather built upon the LTG-BERT model (Samuel et al., 2023) by introducing layer weighting.

3 Experimental Design

3.1 Model Architecture

Like the compound-PCFG system last year and the ELC-BERT model, a custom subword tokenizer was selected for the ELC-ParserBERT model, and it was trained on the provided strict-small data (Georges Gabriel Charpentier and Samuel, 2023; Chen and Portelance, 2023).

The model architecture in this paper uses the ELC-BERT architecture as its backbone (Georges

Gabriel Charpentier and Samuel, 2023) combined with StructFormer's parser network proposed (Shen et al., 2021) with the goal of increased performance from including both the layer weighting and the inductive bias from each, respectively. The architecture, therefore, follows that of the StructFormer but with weighted attention layers from ELC-BERT, as can be seen in Figure 2, where the Parser Network uses a combination of Convolutional layers, Linear layers, and the hyperbolic tangent function to produce the syntactic distances and heights that are used to compute the directed adjacency matrix with probabilities of a token depending on another.

3.2 Data

3.2.1 Training Data

The model uses the provided data from the organizers[2] for the strict-small track, which consists of the following: 8% from the dialogue portion of the British National Corpus (BNC) (Consortium, 2007); 29% from The CHILDES Project's database, a corpora of dialogue concerning child language (MacWhinney, 2000); 26% selected from the standardized Project Gutenberg, a corpus composed of over 50,000 books (Gerlach and Font-Clos, 2018); 20% from Open Subtitles, a corpus of subtitles extracted from movies and television (Lison and Tiedemann, 2016); 15% from nonfiction sections of Simple English Wikipedia[3], an encyclopedia written in plain English to be approachable for English language learners; and 1% from Switchboard, a corpus of dialogues made for dialogue act modeling(Stolcke et al., 2000).

The organizers ran initial preprocessing of the data to ensure that all the data was in plain text, but otherwise left preprocessing open to the contestants. The preprocessing of the training data for this model was largely inherited from the ELC-BERT (Georges Gabriel Charpentier and Samuel, 2023), where standardization is applied to the texts, the texts are compiled, split by line, and segmented into sentences using the Natural Language Toolkit's sentence tokenizer (Bird et al., 2009). After segmentation, the sentences are broken into sequence lengths, encoded by the model's subword tokenizer, and sorted according to their Flesch Reading Ease score (Kincaid et al., 1975) (to allow curriculum learning based upon this metric, if desired).

[2]The training data in its totality with references is available through OSF here: https://osf.io/ad7qg/

[3]https://dumps.wikimedia.org/simplewiki/

3.2.2 Evaluation Data

The model is evaluated on three evaluation benchmarks: BLiMP (as well as BLiMP supplemental) to evaluate the model's knowledge of grammatical phenomena (Warstadt et al., 2020); a selection of tasks that require finetuning from the General Language Understanding Evaluation (GLUE) and its more difficult successor SuperGLUE (Wang et al., 2020, 2018); and the Elements of World Knowledge (EWoK) framework, a benchmark that tests a model's world knowledge by examining the likelihood of context and target pairs across particular domains (Ivanova et al., 2024).

3.3 Experiments

There were two trained models with separate experimental purposes. The first trained model was the ELC-ParserBERT, trained on shuffled data that had a 15% probability of being masked, to be evaluated against the two provided baseline models for the strict-small track. The second model was the curriculum learning ELC-ParserBERT model (referred to hereafter as CL-ELC-ParserBERT), which also had a 15% probability of being masked and was presented in increasing Flesch Reading Ease (Kincaid et al., 1975), but it was compared against the submitted ELC-ParserBERT model in the EWoK evaluation framework. The hyperparameters for both models can be found in Appendix A. The scores were evaluated using the evaluation pipeline provided by the organizers[4].

When evaluating the LTG-BERT baseline model[5] locally, the scores achieved on the EWoK set were found to be different than the scores presented by the organizers. Henceforth, LTG-BERT-A refers to the scores presented by the organizers, and LTG-BERT-B refers to the scores evaluated locally.

Model	BLiMP	Suppl.	EWoK	GLUE	Macroaverage
BabyLlama	69.8	59.5	50.7	63.3	60.8
LTG-BERT-A	60.6	60.8	48.9	60.3	57.7
LTG-BERT-B	60.6	60.8	63.05	60.3	61.2
ELC-ParserBERT	59.6	57.7	63.1	44.5	56.2

Table 1: Model accuracies across different tasks

[4]The pipeline can be found here: https://github.com/babylm/evaluation-pipeline-2024/

[5]The model can be found here https://huggingface.co/babylm/ltgbert-10m-2024

Domains	ELC-ParserBERT	CL-ELC-ParserBERT	ELC-BERT-B
ewok_agent-properties_filtered	0.7376 ± 0.0094	0.7620 ± 0.0091	0.7552 ± 0.0091
ewok_material-dynamics_filtered	0.8104 ± 0.0141	0.8273 ± 0.0136	0.8740 ± 0.0120
ewok_material-properties_filtered	0.6000 ± 0.0377	0.4176 ± 0.0379	0.4647 ± 0.0384
ewok_physical-dynamics_filtered	0.3833 ± 0.0446	0.5083 ± 0.0458	0.3667 ± 0.0442
ewok_physical-interactions_filtered	0.5989 ± 0.0208	0.6025 ± 0.0208	0.6061 ± 0.0207
ewok_physical-relations_filtered	0.8166 ± 0.0135	0.8325 ± 0.0131	0.8166 ± 0.0135
ewok_quantitative-properties_filtered	0.4268 ± 0.0280	0.4013 ± 0.0277	0.5478 ± 0.0281
ewok_social-interactions_filtered	0.5646 ± 0.0290	0.5340 ± 0.0291	0.5374 ± 0.0291
ewok_social-properties_filtered	0.5610 ± 0.0274	0.4573 ± 0.0275	0.4451 ± 0.0275
ewok_social-relations_filtered	0.8068 ± 0.0100	0.7991 ± 0.0102	0.8036 ± 0.0101
ewok_spatial-relations_filtered	0.6347 ± 0.0218	0.6082 ± 0.0221	0.7184 ± 0.0203
ewok total score	0.6310 ± 0.0050	0.6136 ± 0.0050	0.6305 ± 0.0049

Table 2: A breakdown of the accuracies for ELC-ParserBERT, the Learning Curriculum ELC-ParserBERT, and the baseline ELC-BERT performs by each domain in the EWoK evaluation set.

4 Results

The results of the first experiment can be seen in Table 1. Although it performed poorly compared to the baselines in the (Super)GLUE evaluation and had slightly worse BLiMP supplemental scores, ELC-ParserBERT achieved comparable BLiMP scores to the LTG-BERT baselines and had significantly better scores on the EWoK evaluation framework than all other baselines, barring LTG-BERT-B.

Domain Name	p-val	Accuracy
ewok_material-properties_filtered_results	0.0170219	0.14
ewok_quantitative-properties_filtered_results	0.0017656	-0.12
ewok_social-properties_filtered_results	0.0017656	0.12
ewok_material-dynamics_filtered_results	5.97e-05	-0.06
ewok_spatial-relations_filtered_results	3.9e-06	-0.08

Table 3: EWoK domains where ELC-ParserBERT had significant difference in prediction from the LTG-BERT baseline, with the p-value and the change in accuracy relative to ELC-ParserBERT shown.

4.1 EWoK ELC-ParserBERT compared to LTG-BERT-B

ELC-ParserBERT's comparatively strong EWoK predictions prompted further analysis, namely, whether, although LTG-BERT-B and ELC-ParserBert had similar EWoK scores, there was any area where they had statistically significant different predictions. Upon preliminary inspection of the models' EWoK evaluation accuracies broken down by domain, as seen in Table 2, one can already see domains with disparate accuracies despite the close average score. To confirm that these are significant accuracy differences, however, a McNemar test can be constructed for each domain to determine the p-value for the difference in the models' classifications. In Table 3, the domains of the predictions that resulted in a p-value $< .05$ are listed.

These domains, however, can then be further broken down to see which categories within the domains had significant differences by running McNemar tests on the predictions for each group within a domain. In the "material properties" domain, ECL-ParserBERT predicted significantly better for the context type "direct" rather than "indirect," but it cannot be said that this is directly due to the inclusion of the structural bias in the model. It is more likely, however, that there are particular concepts that ELC-ParserBERT understands better or worse than LTG-BERT, although it may be possible that this is indirectly caused by the inclusion of the parser network in the model during pre-training. For instance, ELC-ParserBERT gets all 20 of the instances correct where the context is "cold" or "warm," whereas LTG-BERT only gets 4 of them right. The fourteen more that ELC-ParserBERT predicted correctly all had a "direct" context type, and there is a similar case for the "fragile" and "sturdy" contexts. Moreover, in the "quantitative properties" domain, ELC-ParserBERT actually performs worse in the "direct" context type questions, but this similarly follows from poor performance in particular concepts such as "a lot of" versus "a little."

The categories (concepts, context types, etc.) with significant differences in prediction between ELC-ParserBERT and LTG-BERT within the domains found to have significant differences, as seen

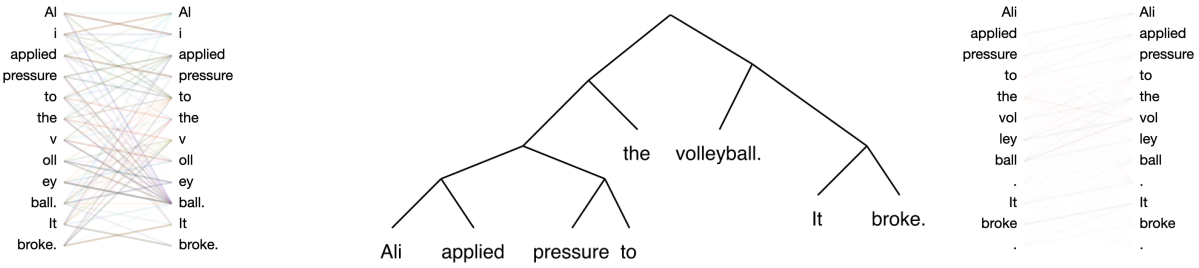

Figure 3: Given one of the contexts for EWoK, this figure shows an attention head of LTG-BERT on the left, the induced tree by ELC-ParserBERT in the middle, and an attention head of ELC-ParserBERT.

Domain Name	p-val	Accuracy
ewok_physical-dynamics_filtered_results	0.028784	-0.12
ewok_social-properties_filtered_results	0.0049673	0.10
ewok_agent-properties_filtered_results	0.0044958	-0.02
ewok_material-properties_filtered_results	0.0011381	0.18

Table 4: EWoK domains where ELC-ParserBERT had significant difference in prediction from CL-ELC-ParserBERT, with the p-value and the change in accuracy relative to ELC-ParserBERT shown.

in Table 3, are enumerated in full in Table 6, located in Appendix B.

4.2 Effectiveness of CL-ELC-ParserBERT

Similarly to LTG-BERT-B, CL-ELC-ParserBERT achieves comparable EWoK scores as seen in Table 2, and, when investigated further, there were four domains with significant difference in prediction between ELC-ParserBERT and CL-ELC-ParserBERT, as can be seen in Table 4. The most notable being the difference in the "material properties" domain, where CL-ELC-ParserBERT has an accuracy 18% smaller than ELC-ParserBERT. Interestingly, again, the concepts of "cold" and "warm" proved difficult for CL-ELC-ParserBERT in the same manner as it did for LTG-BERT-B. CL-ELC-ParserBERT also struggled with the concepts of "heavy" and "light," but it significantly outperformed ELC-ParserBERT when it came to concepts of "sink" and "float," as well as "fall" and "rise."

Although the final scores were close, when breaking down the scores into domains, it's interesting to see how the effects of curriculum learning rather than shuffling, in the context of this training data, result in significantly different predictions for certain domains.

4.3 Attention Comparison

To further examine how the inclusion of the parser network alters the model directly, one can see how the attention differs for a given input, as in Figure 3, by using BertViz (Vig, 2019). The weight of the lines connecting the tokens is based upon the attention between the words. Hence, the dampened weighting of the lines for ELC-ParserBERT shows how the attention is being constrained by the dependency relations produced by the parser network.

5 Conclusions and Future Work

In the context of the BabyLM Challenge 2024, this paper experimented with the ELC-ParserBERT architecture, which is formed by adding the parser network from the StructFormer (Shen et al., 2021) to the ELC-BERT architecture (Georges Gabriel Charpentier and Samuel, 2023). There was no significant improvement found in the BLiMP, BLiMP supplemental, and (Super)GLUE evaluation tasks through the inclusion of the parser network with the training as described. In the EWoK evaluation framework, however, the ELC-ParserBERT architecture showed comparable results to the LTG-BERT-B model and improvement over the other baselines.

This paper also examined the effectiveness of using the Flesch Reading Ease (Kincaid et al., 1975) metric to determine an ordering of the training data for curriculum learning for training the ELC-ParserBERT architecture. The use of this particular learning curriculum on this training data with this architecture did not show any significant improvement generally, but the inclusion or exclusion of this learning curriculum did significantly alter the quality of predictions for certain concepts. Investigating the cause of these particular concept affinities might be the focus for future work.

Future work may also seek to improve upon the ELC-ParserBERT model by ensuring sentences are producing parse trees separately for each sentence in a context window, as is done in GPST (Hu et al., 2024a). Additionally, the ELC-ParserBERT's

largest shortcoming was in the (Super)GLUE evaluation tasks, which employed the default hyperparameters for finetuning set by the organizers, so searching for more optimal hyperparameters may yield overall model improvement.

Acknowledgments

This work was completed in part using the Discovery cluster, supported by Northeastern University's Research Computing team.

References

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

Xuanda Chen and Eva Portelance. 2023. Grammar induction pretraining for language modeling in low resource contexts. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 69–73, Singapore. Association for Computational Linguistics.

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.

BNC Consortium. 2007. British national corpus, XML edition. Oxford Text Archive.

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.

Martin Gerlach and Francesc Font-Clos. 2018. A standardized project gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *CoRR*, abs/1812.08092.

Xiang Hu, Pengyu Ji, Qingyang Zhu, Wei Wu, and Kewei Tu. 2024a. Generative Pretrained Structured Transformers: Unsupervised Syntactic Language Models at Scale. *Preprint*, arXiv:2403.08293.

Xiang Hu, Qingyang Zhu, Kewei Tu, and Wei Wu. 2024b. Augmenting transformers with recursively composed multi-grained representations. In *The Twelfth International Conference on Learning Representations*.

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.

Yoon Kim, Chris Dyer, and Alexander Rush. 2019. Compound Probabilistic Context-Free Grammars for Grammar Induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2369–2385, Florence, Italy. Association for Computational Linguistics.

Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Hongyin Luo, Lan Jiang, Yonatan Belinkov, and James Glass. 2019. Improving Neural Language Models by Segmenting, Attending, and Predicting the Future. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1483–1493, Florence, Italy. Association for Computational Linguistics.

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

Omar Momen, David Arps, and Laura Kallmeyer. 2023. Increasing the performance of cognitively inspired data-efficient language models via implicit structure building. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 327–338, Singapore. 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. *Preprint*, arXiv:2303.09859.

Yikang Shen, Zhouhan Lin, Athul Paul Jacob, Alessandro Sordoni, Aaron Courville, and Yoshua Bengio. 2018. Straight to the Tree: Constituency Parsing with Neural Syntactic Distance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1180, Melbourne, Australia. Association for Computational Linguistics.

Yikang Shen, Yi Tay, Che Zheng, Dara Bahri, Donald Metzler, and Aaron Courville. 2021. StructFormer: Joint Unsupervised Induction of Dependency and Constituency Structure from Masked Language Modeling. *Preprint*, arXiv:2012.00857.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Will we run out of data? limits of llm scaling based on human-generated data. *Preprint*, arXiv:2211.04325.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *Preprint*, arXiv:2303.18223.

A Hyper Parameters

Hyperparameter	Value
Initial learning rate	5e-3
Batch size	256
Steps	13495
Attention probs dropout prob	0.1
Classifier dropout	0.2
Hidden dropout prob	0.1
Hidden size	384
Intermediate size	1024
Layer norm eps	1e-07
Max position embeddings	512
Num attention heads	6
Num hidden layers	12
Vocab size	16384
N parser layers	4
Parser conv size	9

Table 5: Hyperparameters used in the submitted model.

B EWoK Domain and Category Analysis

Domain Name	Category	p-val	Accuracy
material-dynamics	context type - direct	0.01174	-5.8%
material-dynamics	concept - wrinkle	0.00557	-11.7%
material-dynamics	context type - indirect	0.00253	- 6.8%
material-dynamics	concept - stir	0.00074	-8.0%
material-dynamics	target diff - concept swap	5.97e-05	- 6.4%
material-properties	concept - heavy/light	0.04123	37.5%
material-properties	target diff - concept swap	0.01702	13.5%
material-properties	context diff - antonym	0.01219	20.0%
material-properties	concept - cold/warm	0.00051	70.0%
material-properties	context type - direct	1.11e-05	38.5%
quantitative-properties	context type - direct	0.04228	-9.5%
quantitative-properties	concept - a lot of	0.02092	-26.0%
quantitative-properties	context type - indirect	0.01469	-18.5%
quantitative-properties	concept enough/not enough	0.00766	-25.0%
quantitative-properties	target diff - concept swap	0.00177	-12.1%
quantitative-properties	context diff - antonym	0.00103	-15.9%
social-properties	concept - friendly/hostile	0.03888	22.5%
social-properties	context type - indirect	0.01217	14.6%
social-properties	concept - tolerant/bigoted	0.00461	36.0%
social-properties	context diff - antonym	0.00369	12.6%
social-properties	target diff - concept swap	0.00177	11.6%
spatial-relations	context diff - antonym	0.01219	-5.3%
spatial-relations	target diff - concept swap	0.00842	-5.7%
spatial-relations	context diff - variable swap	4.40e-05	-16.4%
spatial-relations	context type - indirect	3.10e-05	-8.5%
spatial-relations	target diff - variable swap	3.04e-05	-15.0%
spatial-relations	concept - above/below	1.19e-07	-14.3%

Table 6: EWoK domains and categories of significant difference between ELC-ParserBERT and LTG-BERT with change in accuracy relative to ELC-ParserBERT.