Syntactic Generalization Under Constraints: Compact Transformer Models and Limited Linguistic Input

Xingming Li
Department of Linguistics and Philology
Uppsala University
Sweden
xingming.li.8883@student.uu.se

Abstract

This study examines the syntactic generalization of two compact transformer models, LTG-BERT-Small and LTG-BERT-XS. When trained on limited child-directed datasets from the BabyLM Challenge 2023 and evaluated with BLiMP and its supplement, both models demonstrate the ability to capture some simple syntactic rules but struggle with complex ones. Compared to human and the pre-trained RoBERTa-Base model, the compact models show some gaps, but still have eye-catching performance. Additionally, LTG-BERT-Small outperforming LTG-BERT-XS demonstrates the effect by parameter capacity. By highlighting the strengths of compact models in handling early-acquired linguistic structures and showing their limitations in capturing hierarchical dependencies, this study offers insights into improving syntactic generalization of compact transformer models under resource constraints.

1 Introduction

Syntactic generalization refers to the capacity to apply learned grammatical rules to new and previously unseen linguistic contexts (Chomsky, 1957). While humans achieve this with limited language input during linguistic development, for language models (LMs) it typically requires extensive training data and computational resources (Devlin et al., 2019). This study investigates whether compact transformer models can replicate human-like syntactic generalization using datasets that simulate children's linguistic exposure (Warstadt et al., 2023a).

Child language acquisition, driven by linguistic and cognitive interaction with their environment, inspired the BabyLM Challenge 2023 datasets, which mimic the linguistic input toddlers and adolescents receive. Recent research on transformer architectures suggest that LMs are capable of capturing syntactic rules beyond surface-level language understanding (Vaswani et al., 2017; Radford et al., 2018). In this study, two compact variants of the LTG-BERT model developed by Samuel et al. (2023), LTG-BERT-Small and LTG-BERT-XS, are trained with child-directed datasets and evaluated to address three key research questions:

- How do certain compact LMs generalize syntactic rules with limited training data?
- How do these models compare to human performance and a large-scale pre-trained transformer model?
- Does parameter capacity influence syntactic generalization in these compact models?

The models are evaluated using the Benchmark of Linguistic Minimal Pairs (BLiMP) and its supplement from the BabyLM Challenge 2023, which assess various syntactic phenomena such as subject-verb agreement and filler-gap dependencies (Warstadt et al., 2020, 2023a). Human performance scores from BLiMP and the pre-trained RoBERTa-Base model serve as skyline comparisons (Liu et al., 2019).

By focusing on compact models, this study stands a chance to provide evidence that certain learning mechanisms in children are analogous to those used by neural networks if these models succeed in generalizing syntax with limited data. This contrasts pre-trained large language models (LLMs) which do not offer the same insights into human language acquisition. The experimental results of this study reveal that while the compact models are on par with or even outperform

RoBERTa-Base on some simpler syntactic tasks, they struggle with more complex phenomena like island effects where humans achieve decent performance. Despite the limitations, the syntactic generalization of compact models still offers valuable insights into the training of LMs under constraints.

2 Related Work

2.1 Child Language Acquisition

The concept of the *poverty of the stimulus*, outlined in Chomsky (1986), argues that the linguistic input available to children is insufficient to explain their rapid acquisition of complex grammatical structures, suggesting the presence of innate linguistic mechanisms. Although the input is limited, imperfect and even sometimes ungrammatical, children quickly acquire various syntactic rules like subject-verb agreement without explicit correction or direct instruction (Mitchell, 2004). This ability to generalize syntactic rules beyond the input supports the idea that children possess innate cognitive mechanisms that guide language acquisition.

Whereas, Tomasello (2000) does not completely second this view, arguing that child language acquisition is more of a result of statistical learning and interaction with the environment. Yang (2016) highlights that capturing hierarchical structures requires not only linguistic exposure, but statistical learning and innate constraints as well. These mechanisms help children abstract grammatical rules and generalize them to novel contexts even without direct examples (Chater and Christiansen, 2003; Howell et al., 2019). Such insights into child language acquisition have inspired this and other LM studies to explore whether computational systems can replicate human-like syntactic generalization by mimicking child-directed input.

2.2 Developmentally Plausible Datasets

A developmentally plausible dataset is typically a collection of data that simulates the type and amount of input available to humans at a certain stage of cognitive or linguistic development. The datasets from the BabyLM Challenge 2023 are appropriate examples since the contents mimic the language input available to children by simplifying grammar and reducing the number of words (Warstadt et al., 2023a). These datasets provide a unique opportunity to explore how models trained with minimal, child-like data perform on syntactic generalization tasks. Comparing LMs trained on such developmentally plausible datasets to patterns in human language acquisition can reveal whether these models replicate human-like syntactic generalization or fall short in capturing key grammatical phenomena.

In addition, unlike children, whose linguistic input during early development is estimated to be tens of millions of words, LLMs such as GPT and BERT are trained on datasets spanning billions of words (Frank, 2023). This difference emphasizes the value of developmentally plausible datasets since they aim to simulate the limited input available to humans. Training on such datasets is a proper starting point to conduct a more direct comparison between LMs and humans to explore how LMs perform under constraints.

2.3 Syntactic Generalization in LMs

When evaluated on syntactic-related tasks, some impressive model performance have already been carried out by transformer models and other LLMs (Devlin, 2018; Brown, 2020). However, it is well known that these LLMs are generally trained with a wide range of corpora (Bommasani et al., 2021). Some research reveal and indicate that LLMs outperform smaller counterparts, but their performance tends to be inconsistent, particularly on hierarchical structures (Hu, 2020; Yedetore et al., 2023). Thus, for current LMs, both the capabilities and limitations to achieve human-like syntactic generalization are underscored in related work (Lin et al., 2021).

3 Data and Methods

3.1 Data

This study utilizes the datasets from the BabyLM Challenge 2023[1] for both training and evaluation. The training data contains two datasets: *Strict* (approximately 100M words) and *Strict-Small* (approximately 10M words), which is also a subset of *Strict* (Warstadt et al., 2023a). As demonstrated in Table 1, these two datasets include diverse sources such as child-directed speech, children's books, and transcribed dialogues in order to simulate the language exposure of children at early linguistic developmental stages. The *Strict-Small* dataset corresponds to the limited linguistic exposure of

[1] https://babylm.github.io/archive_2023.html

Dataset	Domain	*Strict-Small*	*Strict*	Proportion
CHILDES	Child-directed speech	0.44M	4.21M	5%
British National Corpus (Dialogue portion)	Dialogue	0.86M	8.16M	8%
Children's Book Test	Children's books	0.57M	5.55M	6%
Children's Stories Text Corpus	Children's books	0.34M	3.22M	3%
Standardized Project Gutenberg Corpus	Written English	0.99M	9.46M	10%
OpenSubtitles	Movie subtitles	3.09M	31.28M	31%
QCRI Educational Domain Corpus	Educational video subtitles	1.04M	10.24M	11%
Wikipedia	Wikipedia	0.99M	10.08M	10%
Simple Wikipedia	Wikipedia	1.52M	14.66M	15%
Switchboard Dialog Act Corpus	Dialogue	0.12M	1.18M	1%
Total	–	**9.96M**	**98.04M**	**100%**

Table 1: The training datasets of *Strict* and *Strict-Small* of the BabyLM Challenge 2023. The number of words in each included corpus is presented and adapted from Warstadt et al. (2023b)

toddlers (2-3 years old), while the *Strict* dataset reflects the more extensive language input experienced by adolescents (12-13 years old) (Gilkerson et al., 2017).

For evaluation, this study uses the BLiMP benchmark and its supplement in the BabyLM Challenge 2023[2], which test models on a wide range of syntactic phenomena (Warstadt et al., 2020). The evaluation set of BLiMP includes 12 tasks focusing on various grammatical rules and employing minimal-pair sentences to isolate specific linguistic features. For example, in the task of subject-verb agreement, the model must choose between "The grandfathers of Diana drink." (correct) and "The grandfathers of Diana drinks." (incorrect). It is notable that anaphor agreement, argument structure, determiner-noun agreement, subject-verb agreement and irregular forms are likely acquired earlier in development by 12-13 years old (Chien and Wexler, 1990) and (Roeper, 2009). Hence, the performance comparison on these five tasks is more convincing and should be paid extra attention to in this study even though the BLiMP benchmarks are taken from adult annotators. The BLiMP supplement expands the grammatical coverage with 5 additional tasks, such as hypernym and subject-auxiliary inversion, to evaluate broader linguistic competence. For instance, subject-auxiliary inversion task checks if the model can recognize that "Was the book she is reading on the shelf?" is plausible while "Is the book she reading was on the shelf?" is not. Table 2 illustrates the numbers of examples of each task in BLiMP and its supplement used in this study. Together, these benchmarks provide a robust framework for assessing syntactic generalization.

Task	Abbr.	Example
Anaphor Agreement	AA	1956
Argument Structure	AS	8248
Binding	B	6738
Control Raising	CR	4526
Determiner-Noun Agreement	DNA	7542
Ellipsis	E	1732
Filler-Gap	FG	6426
Irregular Forms	IF	1965
Island Effects	IE	2676
NPI Licensing	NL	6586
Quantifiers	Q	3882
Subject-Verb Agreement	SVA	5535
Hypernym	H	860
Q-A Congruence (easy)	QACE	64
Q-A Congruence (tricky)	QACT	165
Subject-Auxiliary Inversion	SAI	4099
Turn-taking	TT	280

Table 2: Number of test examples for each evaluation task of BLiMP and its supplement in BabyLM Challenge 2023. The number of examples after filtering based on the pre-training corpus vocabulary is shown and adapted from Warstadt et al. (2023b). The red phenomena are able to be acquired by 12-13 years old.

3.2 Models

This study evaluates two compact transformer models, LTG-BERT-Small and LTG-BERT-XS, which differ in parameter capacity. Both models share the same architecture with LTG-BERT[3] but vary in hidden size, intermediate size, and the number of attention heads. The configuration files used for training are already developed by Samuel et al. (2023) and the comparison of configuration within these three models is shown in Table 3. LTG-BERT was the winning model in BabyLM Challenge 2023 and is optimized for syn-

[2]Both evaluation data and pipeline can be found in this GitHub repository: https://github.com/babylm/evaluation-pipeline-2023

[3]https://huggingface.co/ltg/ltg-bert-bnc

tactic generalization, which serve as the reason for the model selection. These models are pre-trained during this study on the BabyLM training datasets using the masked language modeling (MLM) objective.

For comparative purposes, the pre-trained RoBERTa-Base model on Hugging Face[4] is included as a skyline, representing near-optimal syntactic generalization achieved by LLMs. Human performance scores reported by Warstadt et al. (2020) are used as another skyline to assess how closely the models replicate human syntactic abilities.

Configuration	XS	Small	Base
Attention Dropout Prob.	0.1	0.1	0.1
Hidden Dropout Prob.	0.1	0.1	0.1
Hidden Size	192	384	768
Intermediate Size	512	1024	2048
Layer Norm Epsilon	1e-07	1e-07	1e-07
Max Pos. Embeddings	512	512	512
# Attention Heads	3	6	12
# Hidden Layers	12	12	12
Position Bucket Size	32	32	32
Torch Data Type	float32	float32	float32
Transformers Version	4.23.1	4.23.1	4.23.1
Vocabulary Size	16384	16384	16384

Table 3: Comparison of model configuration. Adapted from the LTG-BERT GitHub repository with modifications.

3.3 Evaluation

The performance of models is measured using the evaluation pipeline from BabyLM Challenge 2023 with some adaptations. Task accuracy is calculated as the percentage of correct predictions for each task. Moreover, a weighted average accuracy is computed to account for the different number of examples of each task within the evaluation benchmarks. To make sure that the results are representative and accurate enough, each LTG-BERT-Small and LTG-BERT-XS model is evaluated for five rounds in total and their task accuracy scores are averaged to minimize the impact of randomness.

For examining the significance of performance differences between the two compact transformer models, statistical tests are applied to their task accuracy scores. The Shapiro-Wilk test evaluates if the performance difference follows a normal distribution with a null hypothesis that it is normally distributed (Shapiro and Wilk, 1965). If the p-value from the test is greater than a chosen significance level (e.g., 0.05 in this study), the null hypothesis is not rejected, suggesting that the data is likely normal. Then, a paired t-test is run to compare the means of two related samples to detect significant differences. On the other hand, the non-parametric Wilcoxon signed-rank test will be utilized if the data are non-normally distributed in order to examine whether the median differences between the paired samples deviate from zero significantly (Wilcoxon, 1992). These statistical tests, implemented via the `scipy` library, are conducted in order to check if the experimental performance differences between the compact LTG-BERT models are statistically meaningful.

4 Experiments and Results

4.1 Experiments

Hyperparameter	Setup
FF Activation Function	GEGLU
Training Steps	14063
Batch Size	512
Sequence Length	128
Warmup Steps	225 (1.6%)
Initial Learning Rate	0.01
Learning Rate Decay	Cosine
Weight Decay	0.1
Layer Norm ϵ	1e-5
Optimizer	LAMB
LAMB ϵ	1e-6
LAMB β_1	0.9
LAMB β_2	0.98
Gradient Clipping	2.0

Table 4: Pre-training hyperparameters adapted from Samuel et al. (2023) with differences to the original LTG-BERT at numbers of training steps and batch size.

To start the pre-training phase of each LTG-BERT-Small or LTG-BERT-XS model on the *Strict* or *Strict-Small* dataset, the model is initialized with random weights by making use of its corresponding configuration file from the GitHub repository of LTG-BERT[5] and then pre-trained using the MLM objective. The pre-training in this study follows the same pipeline provided in this repository with minimal modifications. The pre-training is conducted on a single A100 GPU on Google Colab[6] using the hyperparameters outlined in Table 4.

[4]https://huggingface.co/FacebookAI/roberta-base

[5]https://github.com/ltgoslo/ltg-bert

[6]https://colab.research.google.com/

Data	Model	AA	AS	B	CR	DNA	E	FG	IF	IE	NL	Q	SVA
Pre-trained	RoBERTa-Base	97.6	83.2	79.6	81.9	96.9	91.9	89.9	95.5	79.2	82.5	71.3	91.7
10M	L-BERT-Small	**91.0**	**70.2**	66.6	**66.5**	**91.3**	**87.7**	**77.6**	**91.0**	**54.6**	**68.9**	**72.5**	**80.8**
10M	L-BERT-XS	88.7	69.4	**67.5**	64.9	89.1	86.1	76.2	90.7	52.5	64.2	70.1	75.4
100M	L-BERT-Small	91.7	**72.0**	**65.9**	**67.6**	**90.5**	**91.1**	**76.7**	93.0	**63.2**	**78.5**	**73.0**	**82.2**
100M	L-BERT-XS	**92.4**	71.5	64.7	64.7	90.0	87.7	75.9	**94.2**	56.2	65.2	72.1	73.5

Table 5: BLiMP results for models trained both on the 100M (*Strict*) and the 10M (*Strict-Small*) BabyLM datasets adapted from Charpentier and Samuel (2023). The bold results represent the better model for the task. The metric used to measure is accuracy. The results are in percentage.

Data	Model	H	QACE	QACT	SAI	TT
Pre-trained	RoBERTa-Base	49.8	93.8	67.1	97.6	72.1
10M	L-BERT-Small	47.3	54.4	47.8	83.4	69.0
10M	L-BERT-XS	47.3	**67.5**	**48.5**	**84.5**	**70.1**
100M	L-BERT-Small	48.1	58.8	**42.0**	82.3	**71.3**
100M	L-BERT-XS	**50.1**	**59.4**	41.7	**85.4**	70.9

Table 6: BLiMP supplement results for models trained both on the 100M (*Strict*) and the 10M (*Strict-Small*) BabyLM datasets adapted from Charpentier and Samuel (2023). The bold results represent the better model for the task. The metric used to measure is accuracy. The results are in percentage.

After pre-training, the models are evaluated on BLiMP and its supplement following the pipeline of the BabyLM Challenge 2023. Each evaluation task is repeated five times with different random seeds to sample 50% of the examples from the whole task set, and the final task accuracy scores are averaged to reduce variability. Task accuracy scores are also subjected to statistical tests in order to assess whether performance differences on BLiMP and its supplement between LTG-BERT-Small and LTG-BERT-XS are statistically significant. These tests are conducted separately for the BLiMP tasks and the BLiMP supplement tasks.

4.2 Results

The experimental results demonstrate the overall syntactic generalization performance of LTG-BERT-Small and LTG-BERT-XS when trained on datasets which mimic the linguistic input to toddlers and adolescents. Both models have the ability to capture some grammatical phenomena well, but have difficulties with others. For specific tasks displayed in Table 5 and 6, both models perform well on tasks that require recognizing simple grammatical structures, such as anaphor agreement (peaking at 92.4% for LTG-BERT-XS trained on *Strict*) and irregular forms (peaking at 94.2% for LTG-BERT-XS trained on *Strict*). However, they struggle on more complex tasks such as filler-gap dependencies and island effects, with accuracy below 80% and 65%, respectively. Tasks in the BLiMP supplement present additional challenges. For instance, performance of most models on hypernym and question-answer congruence (tricky) is below 50%.

Data	Model	BLiMP	B-Supp.
Pre-trained	RoBERTa-Base	86.1	87.8
10M	L-BERT-Small	**75.3**	75.6
10M	L-BERT-XS	73.3	**76.6**
100M	L-BERT-Small	**77.2**	74.9
100M	L-BERT-XS	73.8	**77.5**

Table 7: Weighted average accuracy scores for each model and dataset group, as well as the skyline RoBERTa-base model pre-trained on its full corpora. The bold results represent the better model for the benchmark, except the skyline model. The results are in percentage.

Table 7 demonstrates that the pre-trained RoBERTa-Base model, used as a skyline, significantly outperforms the compact LTG-BERT models. Considering all BLiMP tasks, RoBERTa-Base achieves 86.1% weighted accuracy and 87.8% across the supplement. While RoBERTa-Base performs particularly well on tasks like anaphor agreement and subject-auxiliary inversion, scoring over 97%, the gap between RoBERTa-Base and the compact LTG-BERT models is narrower on some tasks. For example, on ellipsis and turn-taking, LTG-BERT-Small trained on *Strict* is only 0.8% behind RoBERTa-Base.

According to the experimental results, LTG-BERT-Small shows a consistent and obvious advantage against LTG-BERT-XS comparing them on the same training datasets when evaluated on BLiMP. Taking all BLiMP tasks into account, the

	Data		10M	100M
	Model 1		LTG-BERT-Small	LTG-BERT-Small
	Model 2		LTG-BERT-XS	LTG-BERT-XS
BLiMP	Shapiro-Wilk Test	Statistic	0.934	0.831
		p-value	0.423	0.021
	Paired t-test	t-value	**4.024**	-
		p-value	**0.002**	-
	Wilcoxon Signed-Rank Test	W-value	-	**11.000**
		p-value	-	**0.027**
BLiMP Supp.	Shapiro-Wilk Test	Statistic	0.636	0.912
		p-value	0.002	0.479
	Paired t-test	t-value	-	-1.496
		p-value	-	0.209
	Wilcoxon Signed-Rank Test	W-value	1.000	-
		p-value	0.125	-

Table 8: Statistical test results for BLiMP and BLiMP Supplement comparing LTG-BERT-Small and LTG-BERT-XS models. Results include Shapiro-Wilk tests, paired t-tests, and Wilcoxon signed-rank tests. The bold results represent where the differences are significant.

weighted average accuracy of LTG-BERT-Small is 77.2%, while LTG-BERT-XS is 73.8% when trained on *Strict*. The results of the statistical tests presented in Table 8 confirm the significance of such difference with the Wilcoxon signed-rank test yielding a p-value of 0.027 for BLiMP tasks. Similarly, when trained on *Strict-Small*, LTG-BERT-Small also outperforms LTG-BERT-XS on most tasks, and the difference is still significant (p-value = 0.002). However, although the models have differences of performance on BLiMP supplement when trained on either dataset, the statistical tests show that the differences are not significant (p-value > 0.05).

5 Discussion

5.1 Syntactic Generalization with Limited Data

Given the experimental results, it is safe to say that when trained on limited but developmentally plausible datasets, both LTG-BERT-Small and LTG-BERT-XS are still capable of effective syntactic generalization on certain rules. For instance, the models perform well on tasks like determiner-noun agreement and subject-verb agreement and achieve decent accuracy on some of other simple grammatical phenomena. This result is in line with the findings in the prior research which suggests that LMs with compact architectures can still capture basic linguistic structures from relatively small datasets (Manning et al., 2020).

5.2 Comparison of Human and Model Performance

Figure 1 displays accuracy scores on BLiMP tasks by human annotators, pre-trained RoBERTa-Base model and both compact LTG-BERT models trained on *Strict* dataset. The pre-trained RoBERTa-Base model consistently outperforms both LTG-BERT-Small and LTG-BERT-XS across nearly all tasks. This obvious performance gap underscores the advantages of larger parameter capacities and extensive pre-training datasets, which potentially enable RoBERTa-Base to capture syntactic rules. On the other hand, its performance remains below human annotators on complex phenomena. This serves as an evidence that even large-scale pre-training cannot always fully replicate human language acquisition.

Human scores remain the highest across most BLiMP tasks, particularly on linguistically complex phenomena such as binding and island effects. Humans also outperform all evaluated models on tasks like argument structure and irregular forms which are often early-acquired and able to be mastered by 12-13 years old. On the task of argument structure, human score is 6.8% higher than RoBERTa-Base and at least 18% higher than the compact LTG-BERT models. Moreover, humans exceed LTG-BERT-Small and LTG-BERT-XS by over 8.7% on subject-verb agreement. These results align with Yang (2016) and the *poverty of the stimulus* hypothesis which suggest that human language acquisition benefits from innate linguistic mechanisms and constraints beyond language exposure.

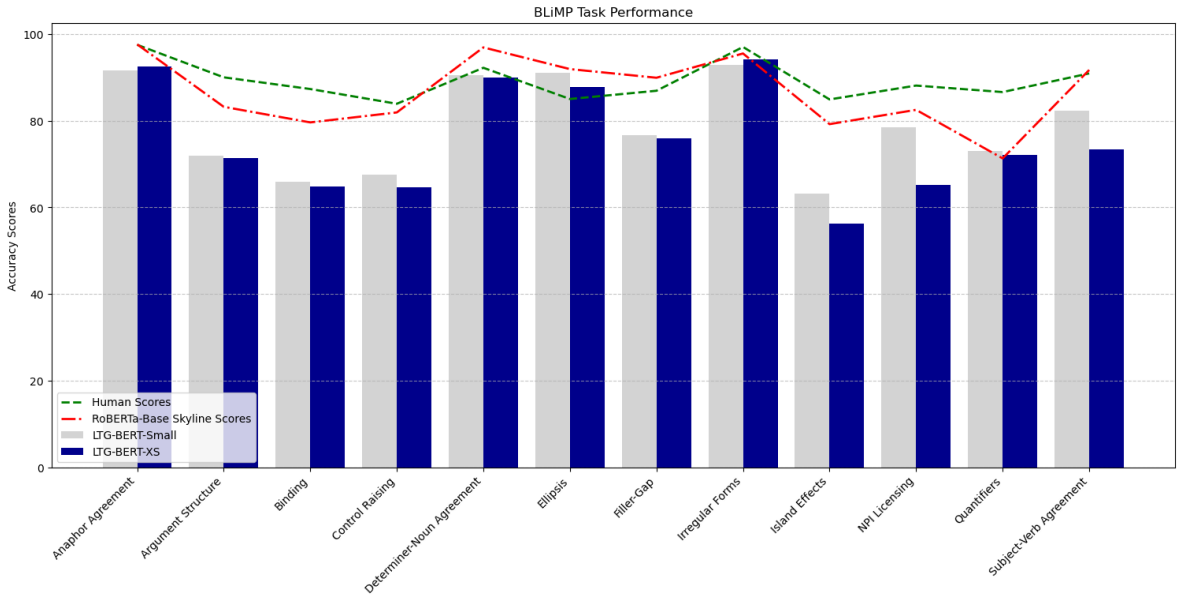

Figure 1: BLiMP task performance comparison of LTG-BERT-Small and LTG-BERT-XS models against human scores and RoBERTa-Base skyline scores. The accuracy scores are in percentage and shown across all linguistic tasks in BLiMP.

It is worth noting that the compact models also produce strong performance of over 90% accuracy on some tasks of early-acquired phenomena, such as determiner-noun agreement and irregular forms, showing their ability to capture simple grammatical rules. Nevertheless, their performance falls way below humans and RoBERTa-Base on tasks related to long-distance dependencies and hierarchical structures, such as island effects and filler-gap dependencies. Theses comparisons underscore not only the challenges LMs face in capturing syntactic rules but a practical direction for model development.

5.3 Impact of Parameter Capacity

The experimental results also demonstrate that the parameter capacity of the model influences syntactic generalization performance tremendously. As confirmed by statistical tests, LTG-BERT-Small outperforms LTG-BERT-XS consistently on BLiMP. These results show that higher parameter capacity contributes to better syntactic generalization performance for LMs. It is even more beneficial for phenomena which require local hierarchical representations like subject-verb agreement. However, when trained on *Strict* dataset and evaluated on simple phenomena such as anaphor agreement and irregular forms, LTG-BERT-XS performs even better than LTG-BERT-Small. These findings suggest that the impact of parameter capacity is likely task-specific and complement one of the earlier research which stresses the importance of adjusting the parameter capacity of the model in order to optimize model performance (Kaplan et al., 2020).

6 Ethical Considerations

This study raises several ethical considerations related to the development and application of LMs. First, the training datasets used are derived from the BabyLM Challenge 2023, designed to simulate child-directed language input. While these datasets provide a good starting point for training under constraints, the data collecting and preprocessing methods may unintentionally oversimplify the complexity of real-world language exposure, including cultural and social aspects.

Second, compact models like the LTG-BERT-Small and LTG-BERT-XS are designed mainly for resource-constrained environments. However, as shown in the experiments, their reduced parameter capacities limit their understanding of complex language patterns. This may potentially lead to errors in sensitive applications such as language-based decision systems. Ensuring transparency of model limitations and appropriate use are critical to mitigating harm. Careful compliance with laws and ethical guidelines is required in order to address these potential risks. Constant review of models can also help avoid related issues.

7 Limitations

This study faces some limitations, particularly in the training phase. The original LTG-BERT-Base model is trained on 128 AMD MI250X GPUs for approximately 8 hours (Samuel et al., 2023). On the contrary, this study only utilizes a single A100 GPU on Google Colab with significantly less computational power. As a result, the LTG-BERT-Small and LTG-BERT-XS models show suboptimal convergence within 14,063 training steps with losses stopping at around 3.5. This likely affects their performance on complex syntactic tasks while further training could yield improvements.

The *Strict* and *Strict-Small* datasets, while developmentally plausible, may not be optimal since the QCRI Educational Domain Corpus in these datasets are substituted by CHILDES in BabyLM Challenge 2024 (Choshen et al., 2024). In addition, since the BabyLM Challenge 2023 also provides multimodal training data, enriching training phase by integrating multimodal context will potentially lead to further advancement of the models' ability to generalize to complex scenarios. Expanding future work to include more varied datasets could also enhance model robustness.

8 Conclusion

This study investigates the syntactic generalization capabilities of LTG-BERT-Small and LTG-BERT-XS, two compact transformer models, and addresses key research questions related to model performance, human-model comparison, and the impact of parameter capacity by training models on child-directed datasets from BabyLM Challenge 2023 and evaluating them with BLiMP benchmarks. Compared to humans and the pre-trained RoBERTa-Base model, the results demonstrate that while both compact models handle basic syntactic rules well, such as determiner-noun agreement and subject-verb agreement, they struggle with more complex phenomena like binding and island effects. In addition, LTG-BERT-Small outperforms LTG-BERT-XS, which also highlights the importance of parameter capacity in syntactic generalization. Alternative pre-training objectives, methods and hybrid architectures could be explored in future research to improve compact models and narrow the performance gap with human language capabilities. Another research direction would be developing age-specific training datasets and benchmarks to refine the human-model comparison in order to help understand how models replicate human-like syntactic generalization under constraints.

References

Rishi Bommasani, Alon Lamm, Edward Wang, et al. 2021. Opportunities and risks in large language models. *arXiv preprint arXiv:2108.07258*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

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

Nick Chater and Morten H. Christiansen. 2003. Connectionism and the poverty of the stimulus. *Cognitive Science*, 27(2):413–440.

Yu-Chin Chien and Kenneth Wexler. 1990. Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language acquisition*, 1(3):225–295.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton.

Noam Chomsky. 1986. *Knowledge of language: Its nature, origin, and use*. Praeger.

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. *arXiv preprint arXiv:2404.06214*.

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

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

Michael C Frank. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*.

Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology*, 26(2):248–265.

Michael Howell, Lilian Sciberras, and John Martin. 2019. Exploring syntactic development through computational modeling. *Cognitive Science*, 43:e12709.

J Hu. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yujia Lin, Yujia Qian, Wei Fu, et al. 2021. Linguistic abilities of transformer-based models: A comprehensive study. *Transactions of the Association for Computational Linguistics*, 9:180–195.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, and Luke Zettlemoyer. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Peter Mitchell. 2004. Acquiring syntactic structures from limited input. *Cognitive Science*, 28(4):613–628.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. In *OpenAI Blog*.

Tom Roeper. 2009. *The prism of grammar: How child language illuminates humanism*. MIT press.

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.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Michael Tomasello. 2000. The item-based nature of children's early syntactic development. *Trends in cognitive sciences*, 4(4):156–163.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

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. *arXiv preprint arXiv:2301.11796*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, et al. 2023b. 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*.

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.

Frank Wilcoxon. 1992. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer.

Charles. Yang. 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. The MIT Press.

Aditya Yedetore, Tal Linzen, Robert Frank, and R Thomas McCoy. 2023. How poor is the stimulus? evaluating hierarchical generalization in neural networks trained on child-directed speech. *arXiv preprint arXiv:2301.11462*.