Mini Minds: Exploring Bebeshka and Zlata Baby Models

Irina Proskurina, Guillaume Metzler, Julien Velcin
Université de Lyon, Lyon 2, ERIC UR 3083, France
Correspondence: Irina.Proskurina@univ-lyon2.fr

Abstract

In this paper, we describe the University of Lyon 2 submission to the STRICT-SMALL track of the BabyLM competition. The shared task is created with an emphasis on small-scale language modelling from scratch on limited-size data and human language acquisition. Dataset released for STRICT-SMALL track has 10M words, which is comparable to children's vocabulary size. We approach the task with an architecture search, minimizing masked language modelling loss on the data of the shared task. Having found an optimal configuration, we introduce two small-size language models (LMs) that were submitted for evaluation, a 4-layer encoder with 8 attention heads and a 6-layer decoder model with 12 heads which we term Bebeshka and Zlata, respectively. Despite being half the scale of the baseline LMs, our proposed models achieve comparable performance. We further explore the applicability of small-scale language models in tasks involving moral judgment, aligning their predictions with human values. These findings highlight the potential of compact LMs in addressing practical language understanding tasks. We make our code and models publicly available.[1]

1 Introduction

LMs accurately encode language-specific phenomena required for natural language understanding and generating coherent continuation of text. LMs gain language understanding about morphosyntax and grammar from large corpora during pre-training. However, they demonstrate partial functional linguistic competence when applying grammatical knowledge to novel expressions at inference time, which is caused by memorising the most occurring linguistic patterns from the training corpus and limited generalization ability of learnt linguistic representations (Wu et al., 2022; Tucker et al., 2022; Mahowald et al., 2023).

Recent pre-training dynamics studies revealed that the performance of LMs can be seen as a function of training corpus vocabulary: (1) grammatical knowledge improves with the expansion of the pre-training data vocabulary (van Schijndel et al., 2019) and (2) small-scale LMs can perform on par with RoBERTa if the vocabulary of used tokenizer is close to the actual human and even child's vocabulary (Liu et al., 2019).

In this paper, we introduce small-scale LMs with an architecture optimized for the STRICT-SMALL track data of BabyLM competition (Warstadt et al., 2023). Our objective is to estimate the general performance and capabilities of shallow LMs in downstream tasks beyond the ones suggested in the evaluation pipeline of shared task. That was achieved through two main contributions.

Contribution 1. We determine an optimal architecture of encoder-based LMs using the Tree-structured Parzen Estimator algorithm and minimal perplexity as a minimizing objective function. Our parameter search results suggest that optimal LMs have a ratio of attention heads to layers around 2, while the ratio of previously tested and existing LMs at their base configuration is equal to one. We introduce new small-scale LMs submitted to the shared task: (*i*) 4-layer encoder Bebeshka[2] and (*ii*) 6-layer decoder Zlata.[3] The parameters of the models are presented in Table 1. Our LMs perform on par with the shared task baselines, while they are half the size of those.

Contribution 2. We investigate the alignment of small-scale LMs predictions with shared human values in the context of moral judgment tasks. We find that shallow LMs, yet trained on limited corpora, perform on par with base LMs in commonsense morality scenarios, and, surprisingly outper-

[1] https://github.com/upunaprosk/small-language-models

[2] A word used to call a baby in a range of South and East Slavic languages.

[3] From Zlato ("Golden sweetheart") used to call babies in West and East Slavic languages.

Proceedings of the 27th Conference on Computational Natural Language Learning: Volume 2: The BabyLM Challenge, pages 58–68
December 6-7, 2023 ©2023 Association for Computational Linguistics

Parameter	RoBERTa	Bebeshka	GPT-2	Zlata
Pre-training objective	MLM	MLM	CLM	CLM
Vocabulary size	50K	8K	50K	30K
#Parameters	125M	16M	345M	66M
Positional embedding type	absolute	rel. key query	absolute	absolute
Maximum sequence length	512	128	1024	1024
(L, A, H, F)	(12, 12, 64, 3072)	(4, 8, 70, 1412)	(24, 16, 64, 4x 1024)	(6, 12, 64, 4x 768)
Activation function	GELU	New GELU	New GELU	GELU
Dropout probability	0.1	0.15	0.1	0.2
Attention dropout	0.1	0.3	0.1	0.2
Processing	1024x V100	4x IPU-M2000	64x V100	4x IPU-M2000
Processing time	1 day	4h	>30 days	6h
Epochs	>40	10	>40	10

Table 1: Model configurations and pre-training details of Bebeshka and Zlata LMs compared to RoBERTa-base and GPT-2 medium. Our LMs have configurations of optimal architecture determined with an architecture search (§3.2). GPT-2 official training information has not been publicly disclosed; we report GPT-2 pre-training hardware details when using model parallelism specified by Shoeybi et al., 2019. We use Graphcore Intelligence Processing Units (IPUs) for pre-training our LMs (Jia et al., 2019 provide a detailed review on IPUs). MLM=Masked Language modelling, CLM=Causal Language modelling, L=Layers, A=Attention heads, H=Hidden size per head, F=Feed-forward (intermediary) layer size.

forming existing baselines in such tasks as virtue and justice assessment. To the best of our knowledge, our work represents one of the earliest attempts to investigate how predictions made by tiny language models trained on a developmentally plausible corpus correlate with human-shared values.

This paper has the following structure. After a short section dedicated to related work (§2), we first describe tokenizer training (§3.1), architecture search results and optimal model selection (§3.2), and the final architecture of the pre-trained LMs (§3.3). Then, we present scores on datasets included in the shared task (§4), and we present ethics evaluation results (§5).

2 Related Work

Recent large LMs found applications in many NLP tasks, such as grammatical correction, text completion, and question answering; yet, their usage is constrained by their computational cost. Previous works reduce the model size and inference time with knowledge distillation, parameter quantization and other compression techniques (Sanh et al., 2019; Yao et al., 2021; Tao et al., 2022). Other studies investigated the relationship between model parameter count and performance. Kaplan et al., 2020 has introduced scaling laws, showing the power-law dependency between perplexity and the model size, as well as between the training loss and dataset size. The paradigm of scaling laws further formed the basis for recent research examining the behaviour of LMs at a small scale (Fedus et al., 2022; Fu et al., 2023). For instance, Puvis de Chavannes et al., 2021 presented results of *Neural Architecture Search* in limited parameter space, suggesting that optimal LMs are smaller than the existing base configurations.

In parallel, there is numerous research focusing on the efficiency of dataset size, vocabulary and representation that can help to reduce computation cost by minimizing the training steps (van Schijndel et al., 2019; Huebner et al., 2021; Schick and Schütze, 2021; Warstadt and Bowman, 2022). van Schijndel et al., 2019 have demonstrated that LMs trained on a small-volume corpus can reach human performance under some grammatical knowledge evaluation scenarios, questioning the necessity of large datasets for pre-training. Huebner et al., 2021 introduced a small encoder-based LM BabyBERTa with 5M parameters and showcased the efficiency of small training data; that work bridged the gap between earlier studies on model size reduction and optimal data size.

The aforementioned related works mainly analyse the difference between compact LMs and their larger counterparts with throughput time measures and performance on GLUE benchmark (Wang et al., 2018). In this paper, we evaluate LMs at a small scale trained on a 10M size dataset of BabyLM shared tasks and try to complement existing research with additional evaluation on moral judgment tasks. The decision to focus on the moral judgment task is driven by recent studies that reveal human-like biases in the moral acceptability judgments made by large language models trained on extensive corpora (Schramowski et al., 2022). This paper complements existing research by conducting a moral judgment evaluation for small language models.

3 Methodology

We follow pre-training tasks of RoBERTa (Liu et al., 2019) and GPT-2 (Radford et al., 2019) and refer to these as the architecture baselines in this section. We train Bebeshka[4] and Zlata[5] with masked language and causal language modelling objectives, respectively, and compare their vocabularies and architectures with the baselines.

3.1 Vocabulary

Training Data We use data provided within the STRICT-SMALL track of the shared task. We report statistics of the training corpus in Table 6 (Appendix A). The transcribed speech, extracted from recordings of casual speech addressed to children and educational movie subtitles, makes up the bulk of the corpus. The average length of the texts is around 30 tokens; considering that and the maximum text length, we lower the maximum sequence length from the base 512 to 128 tokens for the configuration of our LMs.

Input Representation We follow tokenization models of the baselines (GPT-2, RoBERTa) and BabyBERTa (Huebner et al., 2021) and use byte-level Byte-Pair Encoding (BPE) algorithm (Sennrich et al., 2016); that is, a tokenization method based on iterative merging of the most occurring bytes pairs in a further shared vocabulary. For the encoder Bebeshka, we build a case-insensitive vocabulary[6] of size 8K. We find a few mismatches between Bebeshka and RoBERTa tokenization and provide more details in Appendix B. The decoder Zlata has a 30K vocabulary constructed with default parameter settings of Tokenizers trainer;[7] that value also allows for bypassing the inclusion of onomatopoeic words that prevail in some transcribed texts of the shared task data.

3.2 Model Selection

To determine an optimal configuration of encoder LM, we use an Optuna-implemented Bayesian optimization algorithm (Akiba et al., 2019) and tune parameters listed in Table 2 that determine the architecture. The upper bounds of the numerical parameters in a search space are chosen in accordance with the base RoBERTa configuration. We set the lower bounds to 1, ensuring a thorough exploration of architectural variations to find the optimal configuration for the masked language modelling task. Optuna features efficient implementation of optimization algorithms; in our optimization study, we use a standard Tree-structured Parzen Estimator (TPE) algorithm, which uses tree-structured representations and Parzen windows for modelling the probability distributions of hyper-parameters and their density estimation. We use TPE to sample parameter values from the search space and an automated early-stopping based on pruning runs with an intermediary perplexity higher than the median of preceding runs.

We set masked language modelling loss (perplexity) of RoBERTa initialized with the TPE sampled configuration parameters as a minimizing objective function. The perplexity is calculated on the STRICT-SMALL validation set after training the model for 10 epochs on written English texts sample (Gutenberg and Children's Book Test corpora and Wikipedia) from the training BabyLM corpus (see Table 6). We choose a corpus sample to reduce parameter search executing time since dataset size directly impacts an LM training time at each optimization step. We manually found that training on written texts yields a better score. Optimization study with an upper bound of 100 trial runs ran for roughly two days on a single A100 GPU.

Table 2 reports parameter search results for the best and worst runs according to perplexity on the validation dataset.

The **optimal configuration** for encoder LMs can be summarized as follows: (1) the ratio of the number of attention heads to the number of layers fluctuates within the 1.5-2 range, (2) employing relative key query type positional embeddings, (3) the dropping ratio 0.3 for attention probabilities. We further use these three key configuration attributes to initialize Bebeshka. Parameters other than positional embeddings type, dropout ratio and the number of layers/heads vary significantly across the top 10% runs. Precisely, all types of activation functions, except for ReLU, appear evenly in the best range. When it comes to the hidden size per head, it takes values from 65 to 85, with a mean

[4] https://huggingface.co/iproskurina/bebeshka

[5] https://huggingface.co/iproskurina/zlata

[6] We use BPE implementation available under HuggingFace Tokenizers library (Moi and Patry, 2023).

[7] https://github.com/huggingface/tokenizers

Parameter	Search range	10% Best runs Mean	10% Worst runs Mean
Positional embedding type	(rel. key, rel. key query, absolute)	rel. key query	absolute
# Hidden layers	[1-12]	6.2	10.9
# Attention heads	[1-18]	11.9	7.1
Hidden size per head	[1-100]	81.6	64.1
Feed-forward layer size	[1-3072]	1446.3	2034.5
Activation function	(New GELU, GELU, SiLU, ReLU)	New GELU	ReLU
Dropout probability	[0.1-1.0]	0.16	0.63
Attention dropout	[0.1-1.0]	0.33	0.70
Avg. perplexity	-	24.53	992.27

Table 2: Parameter search space of Optuna study for pre-training encoder LMs on Strict-Small corpus and mean parameter values across 10 best and worst runs sorted by the perplexity. For non-numerical parameters, we report the most common parameter values among study runs.

Model	Loss		Run time	
	Val	Test	Val	Test
	MLM			
RoBERTa (125M)	3.72	4.42	1519	1592
Bebeshka (16M)	**3.54**	**4.30**	**485**	**649**
	CLM			
OPT (125M)	7.11	7.10	1493	1567
Zlata (66M)	4.64	4.69	831	869

Table 3: Pre-training objective loss on validation and test data of Bebeshka and Zlata compared to baseline models and average run time in seconds. We run an evaluation of all LMs on the same V100 GPU and use Hugging Face Trainer API for calculating the scores. The best score is in bold, and the second-best score is underlined.

of 81.6. We also observe a notable deviation of intermediary size from the mean value. Altogether our results show that the best-performing encoder LMs are smaller than the base configuration of RoBERTa, which aligns with Puvis de Chavannes et al., 2021.

3.3 Model Pre-training

We train our models on 4 Graphcore IPUs with two encoder layers trained on each with mixed precision[8] and use Strict-Small training split. Table 1 shows the configuration settings of our LMs.

Bebeshka The 16M parameters model is based on RoBERTa architecture with determined optimal layer sizes (§3.2). We train Bebeshka on the 10M training corpus of the shared task. We decrease the probability for selecting masked tokens from standard 15% to 13.5%, which is one of the equivalents to set RoBERTa unmasking probability to 0 discussed by Huebner et al., 2021.

Zlata That decoder LM is a light 66M version of GPT-2 with 6 layers trained for 10 epochs on the training Strict-Small data. Motivated by the configuration of the best encoder LM, we use the ratio of attention heads to decoder layers equal to 2. We explain parameter choice in Appendix C.

4 Experiments Results

In this section, we report the results submitted for the BabyLM shared task. LMs discussed in this section are pre-trained on the shared task data, including the baselines. We use baselines that were created with existing tokenizers and released by the organizers of the BabyLM competition.[9]

4.1 Pre-training Objective Loss

We present the evaluation results of our LMs in Table 3, where we compare their performance against the shared task baselines and evaluation runtime. While the baselines were trained for 20 epochs, we can observe competitive results by pre-training our small-scale models for ten epochs. One of the main advantages of the introduced models lies in their compact size, which makes them more efficient at inference time, even though they do not outperform the baselines by a large margin, which can be seen from the average run time.

4.2 Linguistic Minimal Pairs

Figure 1 depicts the evaluation results of our LMs on the BLiMP dataset (Warstadt et al., 2020a) in a zero-shot setting. The goal of this evaluation benchmark is to assess a model's ability to distinguish between grammatically acceptable and unacceptable sentences without specific fine-tuning on the task. The dataset consists of minimal pairs annotated

[8] https://www.graphcore.ai/products/ipu

[9] We also report scores for the version of the model trained with full precision weights, which we dub Bebeshka-2. However, we do not discuss those since they were submitted after the leaderboard release.

Model	CoLA MCC	SST-2 Acc.	MRPC F1	QQP F1	MNLI Acc.	MNLI$_{\text{mm}}$ Acc.	QNLI Acc.	RTE Acc.	BoolQ Acc.	MultiRC Acc.	WSC Acc.
OPT	15.2	81.9	72.5	60.4	57.6	60.0	61.5	60.0	63.3	55.2	60.2
RoBERTa	**25.8**	**87.0**	79.2	73.7	**73.2**	**74.0**	**77.0**	**61.6**	**66.3**	**61.4**	61.4
T5	11.3	78.1	**80.5**	66.2	48.0	50.3	62.0	49.4	66.0	47.1	61.4
Bebeshka	0.11	81.3	73.5	66.4	58.7	62.0	59.0	45.4	63.9	48.7	61.4
Zlata	0.05	81.7	77.6	65.9	61.9	63.9	61.7	56.6	65.3	53.8	**61.5**
Bebeshka-2	24.5	83.5	77.7	**77.3**	65.4	66.9	64.0	56.6	60.2	46.9	61.4

Table 4: Evaluation results on GLUE and SuperGLUE (BoolQ, MultiRC, WSC) benchmark datasets. We report metrics suggested in the shared task evaluation pipeline and baselines. The best score is in bold, and the second-best score is underlined.

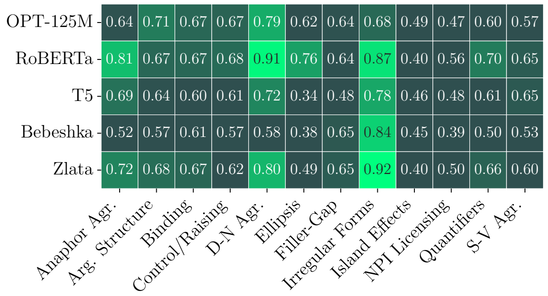

Figure 1: Accuracy on BLiMP tasks of our LMs with RoBERTa-base, OPT-125M, and T5-base baselines. The lighter colours correspond to greater accuracy and, hence, better scores. Morphology: *Anaphor Agr.*, *D-N Agr.*, *Irregular Forms*, *S-V Agr.*. Semantics: *NPI Licensing*, *Quantifiers*. Syntax-Semantics:*Binding*, *Control/Raising*. The rest phenomena correspond to the Syntax category.

with a grammatical phenomenon. We report detailed LMs accuracy scores across various BLiMP tasks in Table 7 (Appendix D). The general trend is that LMs trained on BabyLM data perform well on minimal pairs with morphological tasks, such as *Irregular Forms* and *Determiner-Noun Agreement*.

Zlata achieves the best accuracy (92.1%) on *Irregular Forms* and outperforms OPT-125M baseline on some morphological tasks (*Anaphor Agreement*, *Subject-Verb Agreement*), minimal pairs with a violation in phrasal movements (*Filler Gap*) and other tasks, such as *NPI Licensing*. Bebeshka achieves the second-best accuracy (64.7%) on *Filler Gap* minimal pairs and distinguishes sentences with syntactic errors in pronoun and its antecedent relationship or syntactic islands (*Binding*, *Island Effects*). The results show that LMs trained on the BabyLM corpus have syntactic and morphology understanding which influences their behaviour on downstream tasks discussed next.

4.3 GLUE

Table 4 shows results of fine-tuned LMs evaluation on a variety of tasks present in GLUE and SuperGLUE benchmarks.[10] Submitted to the shared task, Bebeshka and Zlata were fine-tuned for ten epochs on most of the tasks (see Appendix C for more detail). The overall trend is that the introduced small-scale encoder Bebeshka and decoder Zlata demonstrate scores comparable with large baseline LMs on downstream tasks. That highlights that LMs at a small scale can quickly adapt to the fine-tuning task, though may achieve lower performance in a zero-shot evaluation on BliMP. When comparing decoder LMs, we observe that the introduced Zlata outperforms OPT-baseline on paraphrase detection (MRPC & QQP), entailment/contradiction detection (MNLI), and question answering (BoolQ) downstream tasks. As for the encoder LMs, the encoder Bebeshka has moderate scores compared to RoBERTa, which, in general, achieves the best scores on GLUE. However, Bebeshka outperforms OPT-125M baseline on QQP and MRPC tasks with F1 scores of 73.5% and 66.4%, respectively.

The most difficult task for shallow LMs seems to be Recognizing Textual Entailment (RTE). We suppose that LMs trained on STRICT-SMALL corpus with an average length of 28.65 tokens (Table 6, Appendix D) or restricted to the 128 maximum sequence length, can perform well on datasets with short sequences and contexts, which can explain lower results on some fine-tuned tasks; another issue can be the fine-tuning hyper-parameters search: perhaps, shallow LMs require more epochs to improve the submitted scores.

[10] Provided datasets within the shared task were filtered according to the vocabulary of BabyLM STRICT-SMALL corpus.

Model	Justice	Deontology	Virtue	Utilitarianism	Commonsense
RoBERTa-large (355M)	<u>56.7</u>	<u>60.3</u>	53.0	**79.5**	**90.4**
GPT-3 few-shot (175B)	15.2	15.9	18.2	<u>73.7</u>	<u>73.3</u>
Bebeshka (16M)	**64.6**	**71.4**	**74.1**	69.0	-
Zlata few-shot (66M)	50.7	49.6	<u>72.0</u>	50.3	53.3

Table 5: Accuracy scores on ETHICS benchmark. LMs trained on STRICT-SMALL corpus reach results close to the large model baselines reported by Hendrycks et al., 2020. We do not report results for the fine-tuning tasks which require the maximum sequence length exceeding the one of an LM. The best score is in bold, and the second-best score is underlined.

4.4 Mixed Signals Generalization

The MSGS dataset introduced by (Warstadt et al., 2020b) comprises 20 binary classification tasks and is used to test whether a LM has a preference for linguistic or surface generalizations. The evaluation pipeline of the shared task includes 11 MSGS tasks; we report obtained accuracy scores for the fine-tuned LMs in Table 8 (Appendix D). The Matthew's Correlation Coefficient (MCC; Matthews, 1975) scores suggest that all LMs fine-tuned in a controlled setting show better results (>0.9) than those fine-tuned in an ambiguous scenario, with the only exception for *Control Raising* category; the highest scores are achieved on *Lexical content* and *Relative position* tasks. *Lexical Content* is a task of classifying sentences with "the" (*the mouse* vs *a mouse*) when *Relative Position* is a task of determining whether "the" precedes "a" in a sentence. Decoder LMs perform similarly on MSGS tasks chosen for the BabyLM competition, excluding *Syntactic Category-Lexical Content* (SC-LC) classification task, where SC is a task of detecting sentences with adjectives. A decoder LM Zlata seems to adopt surface generalization during fine-tuning on unambiguous data (SC-LC), whereby the baseline model OPT learns to represent linguistic features. Bebeshka behaves likewise on the *Syntactic Category* task and reaches scores close to RoBERTa on *Lexical Content* and *Main Verb* classification problems, suggesting that Bebeshka tends to encode surface features.

4.5 Age of Acquisition

Portelance et al., 2023 introduced a method for measuring the age-of-acquisition in LMs compared to the actual age-of-acquisition by English American children on words set from the CHILDES corpus. Table 9 (Appendix D) illustrates that deviation measured in months for the introduced and baseline LMs. The models Zlata and Bebeshka demonstrate comparable scores to the baselines.

5 Moral Judgments

In this section, we present the results of additional experiments on moral judgements that we conduct outside of the main shared task evaluation.

We evaluate small-scale LM's understanding of fundamental moral principles in various scenarios covered by ETHICS benchmark (Hendrycks et al., 2020). The benchmark consists of 5 morality judgment tasks, including reasonable and fair justice, virtue responses, permitted behaviour depending on context-specified constraints (deontology ethics), pleasant scenario choice (utilitarianism ethics), and commonsense morality. We grid search hyper-parameters for our LMs and use test splits for further evaluation. We fine-tune Bebeshka for ten epochs on each of the tasks and evaluate Zlata in a few-shot setting (see more details in Appendix C). Table 5 outlines the moral judgements classification results. Our small LMs generally outperform existing baselines with respect to accuracy scores on sentence-level tasks, and the best results are achieved on *Virtue* moral judgements.

We suggest that the efficiency of small LMs in these tasks can be explained by some properties of pre-training data, such as lower mean sequence length, transcribed speech prevalence with single-word reactions or responses, children-directed speech, and imperatives. For example, *Virtue* task is a collection of scenario-trait pairs, such as *"Jordan will never do harm to his friends. <sep> caring"*, which have a structure similar to one-word responses in transcribed dialogues.

6 Conclusion and Future Work

In this paper, we present our results for the STRICT-SMALL track of the BabyLM competition. Our submission to the shared task consists of two LMs, namely encoder Bebeshka and decoder Zlata. We first search for an optimal architecture, minimizing perplexity on the released training corpus, and find that the best models have around 6 encoder layers on average, down from 12 layers of existing base models, and have twice as many attention heads. When the number of encoder layers fluctuates among the best models, we find that they all have an attention-heads-to-layers ratio of two, which we further use for building our LMs. Our final LMs, which are scaled-down versions of RoBERTa and GPT-2 with a total of 16M and 66M parameters, perform better than the baseline LMs on development and test BabyLM corpora. Zero-shot evaluation results suggest that our shallow LMs have some basic grammatical knowledge of language syntax and morphology. The introduced LMs also perform better than OPT model on several downstream tasks when having 2 times fewer parameters. We also observe a good performance of our small LMs in a range of ethics judgment tasks, showing that their vocabulary and after-training knowledge can positively contribute to the morality assessment of the described scenarios. These results can serve as baselines for the evaluation of ethical judgment capabilities in small language models. The achieved scores may be attributed to the interplay between ethical and linguistic rules, particularly in encoding action verbs used to describe moral and immoral behaviour. This aspect can be further explored by examining the usage of verbs in various syntactic contexts within the BabyLM corpus and their encoding by trained language models.

In our future work, we plan to determine more capabilities of small LMs, trained on small-size corpora, such as short stories data containing words only 4-year-old children can understand (Eldan and Li, 2023). We also plan to extend our experiments with an analysis of fine-tuning dynamics to investigate how small models adapt to the tasks.

Limitations

Despite achieving good performance on BabyLM test data, our approach has some limitations. We use a variant of Bayesian optimization (TPE algorithm, §3.2) to find an optimal range of parameters that we further use for building our LMs. We predefine constraints for parameters (Table 2) that narrows down the search space and can influence further parameter distributions built with Parzen (kernel density) estimators and, thus, future candidate selection. Future work can benefit from both expanded search space and parameter limits range. The architecture of our small language models, including the number of layers, heads, and hidden layer size, can serve as a minimum lower bound for the parameter search space.

Acknowledgements

This work was funded by the ANR project Diké (grant number ANR-21-CE23-0026-02).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1):5232–5270.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning AI with shared human values. *arXiv preprint arXiv:2008.02275*.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Zhe Jia, Blake Tillman, Marco Maggioni, and Daniele Paolo Scarpazza. 2019. Dissecting the graphcore ipu architecture via microbenchmarking. *arXiv preprint arXiv:1912.03413*.

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*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models: a cognitive perspective. *arXiv preprint arXiv:2301.06627*.

Brian W Matthews. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

Anthony Moi and Nicolas Patry. 2023. HuggingFace's Tokenizers.

Eva Portelance, Yuguang Duan, Michael C. Frank, and Gary Lupyan. 2023. Predicting age of acquisition for children's early vocabulary in five languages using language model surprisal.

Lucas Høyberg Puvis de Chavannes, Mads Guldborg Kjeldgaard Kongsbak, Timmie Rantzau, and Leon Derczynski. 2021. Hyperparameter power impact in transformer language model training. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, pages 96–118, Virtual. Association for Computational Linguistics.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4821–4836, Dublin, Ireland. Association for Computational Linguistics.

Mycal Tucker, Tiwalayo Eisape, Peng Qian, Roger Levy, and Julie Shah. 2022. When does syntax mediate neural language model performance? evidence from dropout probes. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5393–5408, Seattle, United States. Association for Computational Linguistics.

Marten van Schijndel, Aaron Mueller, and Tal Linzen. 2019. Quantity doesn't buy quality syntax with neural language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5831–5837, Hong Kong, China. Association for Computational Linguistics.

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 and Samuel R Bowman. 2022. What artificial neural networks can tell us about human language acquisition. *Algebraic Structures in Natural Language*, pages 17–60.

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

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

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Zhengxuan Wu, Atticus Geiger, Joshua Rozner, Elisa Kreiss, Hanson Lu, Thomas Icard, Christopher Potts, and Noah Goodman. 2022. Causal distillation for language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4295, Seattle, United States. Association for Computational Linguistics.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 460–470, Online. Association for Computational Linguistics.

A Experimental Framework

Dataset	# Sentences	Avg. length*	Questions (Proportion)	Proportion
CHILDES	64258	7.17	39%	5%
British National Corpus (BNC)	66100	16.06	17%	8%
Children's Book Test	25946	25.49	3%	6%
Children's Stories Text Corpus	5569	60.58	1%	3%
Standardized Project Gutenberg Corpus	90402	16.22	0%	10%
OpenSubtitles	417984	9.94	17%	31%
QCRI Educational Domain Corpus	91904	16.38	0%	11%
Wikipedia	40876	51.28	0%	10%
Simple Wikipedia	9938	14.57	6%	15%
Switchboard Dialog Act Corpus	5569	60.58	0%	1%
Total	832274	28.65	13.1%	100%

Table 6: Statistics of the training corpus offered in the STRICT-SMALL track of BabyLM competition. *= Average tokenized text length.

B Tokenization Tests

We compare the tokenization of Bebeshka and RoBERTa on the corpus of STRICT-SMALL track and find that the tokenization coincides on 87% of the sequences. We manually analyse a random sample of 100 non-matching tokenization cases and find that those fall on transcribed speech sentences with no more than three words or include two words missing in RoBERTa vocabulary but processed as a whole word by Bebeshka LM (*sweetie* and *duke*). We also found that the RoBERTa tokenizer splits non-capitalised first names or other terms used for addressing (*th-omas*, *m-ister*, *mom-my*) opposed to Bebeshka.

C Training Details

C.1 Pre-training parameters

We experimented with the same configuration for our decoder LM Zlata as we used for Bebeshka, including 4 layers and the same type of positional embeddings; however, that always resulted in gradients underflow and that loss was not decreasing. We manually found the 6-layer and absolute positional embedding configurations by increasing and traversing values of the parameters that were grid searched for Bebeshka (Table 2). We pre-train our LMs using 4x IPUs freely available in Paperspace[11] and use IPU Trainer API. We use auto-loss scaling with an initial value of 16384 and half-precision for training our LMs. Training with IPUs requires specifying IPU configuration, containing instructions for mapping layers between the devices; for Bebeshka, we use one layer per IPU, and for Zlata, we use that parameter equal to 2. For both LMs, we use per-device training batch size equal to 1 and gradient accumulation steps equal to 64. Each batch consists of 1,000 concatenated data examples from the training corpus. The time for the computational graph construction took under 10 minutes for both training both LMs.

C.2 Fine-tuning parameters

BabyLM Evaluation For Bebeshka fine-tuning, we use parameters used by default in the evaluation pipeline of the competition, that is, learning rate equal to 5e-5, batch size equal to 64, and maximum epochs equal to 10. For Zlata fine-tuning, we use the learning rate equal to 1e-4 and fine-tune the tasks for 5 epochs. That allowed us to reduce fine-tuning time. Note that the performance of our LMs can be improved upon the submitted results if grid search the optimal hyper-parameters.

Moral Judgement We use a weighted loss for fine-tuning Bebeshka and grid search optimal parameters using an official implementation by the authors of the dataset.[12] For our GPT-2 based model Zlata, we use an existing evaluation harness benchmark in the k-shot setting with k equal to 15.[13]

[11] https://www.paperspace.com

[12] https://github.com/hendrycks/ethics

[13] https://github.com/EleutherAI/lm-evaluation-harness/

D Evaluation Results

Model	Anaphor Agr.	Arg. Structure	Binding	Control/Raising	D-N Agr.	Ellipsis	Filler-Gap	Irregular Forms	Island Effects	NPI Licensing	Quantifiers	S-V Agr.
OPT-125M	63.8	**70.6**	67.1	66.5	78.5	62.0	63.8	67.5	**48.6**	46.7	59.6	56.9
RoBERTa-base	**81.5**	67.1	67.3	**67.9**	**90.8**	**76.4**	63.5	87.4	39.9	**55.9**	**70.5**	**65.4**
T5-base	68.9	63.8	60.4	60.9	72.2	34.4	48.2	77.6	45.6	47.8	61.2	65.0
Bebeshka	52.0	57.3	61.5	56.8	58.0	37.9	64.7	84.5	44.8	39.2	49.7	53.2
Zlata	72.0	68.1	66.9	61.7	80.0	48.6	**65.4**	92.1	40.3	50.4	66.4	60.3
Bebeshka-2	77.7	60.2	**68.0**	56.2	87.4	68.8	64.7	**92.8**	37.0	45.1	70.2	60.5

Table 7: Model evaluation results on BLiMP dataset. The scores show the model's accuracy in distinguishing between the grammatical and ungrammatical sentences within each minimal pair. The best score is in bold, and the second-best score is underlined.

Model	**CR**	**LC**	**MV**	**RP**	**SC**	**CR LC**	**CR RTP**	**MV LC**	**MV RTP**	**SC LC**	**SC RP**
	Control					**Ambiguous**					
OPT	**50.8**	53.6	**99.5**	**99.9**	77.2	**0.4**	-70.3	-72.1	-77.6	13.8	-68.9
RoBERTa	43.1	**100.0**	97.7	76.7	**86.2**	-28.3	-77.7	-99.3	-79.4	**16.3**	-45.0
T5	21.1	**100.0**	33.4	82.5	77.6	-78.3	**-62.0**	-100.0	-79.7	-25.3	-39.4
Bebeshka	13.0	**100.0**	97.0	72.0	41.0	-95.0	-63.0	-100.0	-66.0	-58.0	-62.0
Zlata	37.0	79.0	90.0	87.0	64.0	-9.0	-85.0	**-70.0**	-94.0	-58.0	**-39.0**
Bebeshka-2	49.4	**100.0**	98.2	88.3	61.5	-28.9	-80.4	-100.0	**-40.8**	-57.2	-46.4

Table 8: Model evaluation results: Matthews Correlation Coefficient (MCC) on the synthetic MSGS dataset, multiplied by 100. CR=*Control Raising*, LC=*Lexical Content*, MV=*Main Verb*, RP=*Relative Position*, SC=*Syntactic Category*, RTP=*Relative Token Position*. Control columns correspond to the control experiments when an LM is trained to classify sentences with certain linguistic and surface features. Ambiguous correspond to the experiments when an LM is tested on a single-feature dataset (for example, LC) after training on a set with labels consistent across both linguistic and surface features (SC LC). The highest score is in bold, and the second-highest score is underlined.

Model	**Overall (591 words)**	**Nouns (322)**	**Predicates (167)**	**Function words (102)**
OPT-125M	2.03	1.98	1.81	2.57
RoBERTa-base	2.06	1.99	1.85	2.65
T5-base	2.04	1.97	1.82	2.64
Bebeshka	2.06	1.98	1.84	2.66
Zlata	2.07	1.99	1.83	2.67

Table 9: Age-of-acquisition (AoA) predictions on child-directed utterances from CHILDES data. The scores are Mean Absolute Deviation scores in months between the actual average AoA of the words by American English-speaking children and model predicted AoA, measured as a likelihood of the words' usage across all the contexts (surprisal scores). The lower the MAD scores, the better. Top-5 words with the highest surprisal scores for LMs: Zlata: *snowsuit*, *applesauce*, *lawn mower*, *sprinkler*, *tricycle*; Bebeshka: *snowsuit*, *hen*, *turkey*, *belt*, *lamb*.