BabyLlama-2: Ensemble-Distilled Models Consistently Outperform Teachers With Limited Data

Jean-Loup Tastet
University of Copenhagen
Department of Computer Science
Copenhagen, Denmark
jeta@di.ku.dk

Inar Timiryasov
University of Copenhagen
Niels Bohr Institute
Copenhagen, Denmark
inar.timiryasov@nbi.ku.dk

Abstract

We present BabyLlama-2, a 345 million parameter model distillation-pretrained from two teachers on a 10 million word corpus for the BabyLM competition. On the BLiMP and SuperGLUE benchmarks, BabyLlama-2 outperforms baselines trained on both 10 and 100 million word datasets with the same data mix, as well as its teacher models. Through an extensive hyperparameter sweep, we demonstrate that the advantages of distillation cannot be attributed to suboptimal hyperparameter selection of the teachers. Our findings underscore the need for further investigation into distillation techniques, particularly in data-limited settings.

1 Introduction

With frontier model training runs using beyond 10^{25} FLOPs (Dubey et al., 2024), training efficiency has become a billion-dollar question. Humans are vastly more sample efficient than current Large Language Models (LLMs). For example, a typical 13-year-old child has been exposed to less than 100 million words (extrapolating from Gilkerson et al. (2017)), whereas Llama-3.1 has been trained on 15.6 trillion text tokens. The goal of the BabyLM Challenge (Choshen et al., 2024) is to optimize pretraining given dataset limitations inspired by human development.

In this work, we present our contribution to the BabyLM challenge (Strict-Small Track), with the following main results:

- BabyLlama-2 model: This 345M parameter decoder-only model[1], distillation-pretrained on 9.5M words, outperforms baseline models trained on both 10M and 100M words (using the same data mix). It also surpasses similar models pretrained using conventional methods.
- Extensive hyperparameter sweep: We have conducted a comprehensive hyperparameter optimization and demonstrated that distillation pretraining consistently outperforms the best models from the sweep.
- Correlation between test loss and performance: As a byproduct of our sweep, we have identified a correlation between zero-shot performance on the BLiMP task and the model's test loss.

The success of distillation pretraining, i.e. pretraining from scratch with distillation loss, in our experiments highlights its potential as a powerful technique for improving model performance, especially in data-limited settings. While our findings are promising, they also raise intriguing questions about the nature of knowledge distillation and its interaction with pretraining objectives. Further investigation into these areas could yield valuable insights for the development of more sample-efficient language models.

2 Related Work

The first edition of the BabyLM challenge, which aims to optimize language model pretraining under data constraints inspired by human language acquisition, prompted numerous works on sample-efficient pretraining. For a detailed summary of all contributions, see the review by Warstadt et al. (2023).

Outside the BabyLM context, relatively few works address training on limited language datasets. Notable exceptions include Muennighoff et al. (2023), who studied the scaling of data-constrained

[1] It is worth noting that encoder models are better suited for the evaluation tasks of the challenge than decoder ones. In last year's evaluation (Warstadt et al., 2023), the 125M parameter RoBERTa-base (Liu et al., 2019) performed on par with the 70B parameter Llama-2 (Touvron et al., 2023b). However, our focus throughout this paper shall be on generative, decoder models.

The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning, pages 292–301
November 15-16, 2024 ©2024 Association for Computational Linguistics

LLMs. Their main finding is that training for more than 4 epochs leads to diminishing returns. Luukkonen et al. (2023) trained FinGPT on more than 30B tokens in Finnish language. Although resource-constrained, this dataset is significantly larger than that of the BabyLM Challenge. A sample-efficient modification of BERT architecture was proposed by Samuel et al. (2023), with a model trained on a 100M word dataset from the British National Corpus outperforming the original BERT model.

The existing literature on training small models often focuses on models deployable on edge devices, such as MobileLLM (Liu et al., 2024). However, these works typically concentrate on deployment efficiency rather than sample efficiency.

Knowledge distillation has recently attracted significant attention, primarily for deployment efficiency reasons (see Xu et al. (2024) for a systematic review). Typically, this involves using large frontier models as teachers to train smaller student models. In contrast, BabyLlama-2 utilizes distillation for sample-efficient pretraining, using similar-sized teacher models trained on the same limited dataset.

A similar phenomenon, where a student model outperforms its teachers when distilled from models with identical architecture and trained on the same dataset, was observed in “Born-Again Neural Networks” (Furlanello et al., 2018). However, this work did not focus on the data-limited regime and it used LSTM variants (instead of transformers) for language modeling.

3 Background

Knowledge distillation, introduced by Hinton et al. (2015), is a technique for transferring knowledge from a “teacher” model to a “student” model. The core idea is to train the student to mimic the logit distribution (soft targets) produced by the teacher, rather than just the hard labels of the training data. The distillation loss combines the standard cross-entropy loss with the soft target loss:

$$\mathcal{L}_{\text{distill}}(y, z_s, z_t) = \alpha\, \mathcal{L}_{\text{CE}}(y, \sigma(z_s)) + (1-\alpha)\, T^2\, D_{\text{KL}}\left(\sigma(z_t/T)\,||\,\sigma(z_s/T)\right) \quad (1)$$

where α balances the usual cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the soft targets loss, T is the temperature parameter that softens the probability distributions, z_s and z_t are respectively the logits of the student and teacher models, σ is the softmax function, and D_{KL} denotes the Kullback-Leibler divergence. In our implementation, we use the averaged logits of an ensemble of teacher models as z_t. Moreover, unlike typical applications, our student and teacher models are of the same size.

4 Model

Architecture. Previous experiments have shown that the Llama architecture (Touvron et al., 2023a), featuring RoPE and a SwiGLU non-linearity, requires fewer epochs to reach minimal loss compared to GPT-2 or GPT-J architectures (Timiryasov, 2023). After training a family of Llama models ranging from 16M to 728M parameters, we converged on a specific 345M model architecture suggested in MobileLLM (Liu et al., 2024) and also used in SmolLM (Allal et al., 2023), whose hyperparameters are listed in table 1. This design incorporates Grouped-Query Attention (GQA) and prioritizes depth over width. Some details of our model selection are listed in appendix B.

Hyperparameter	Value
Vocabulary size	16,000
Number of layers	32
Number of heads	15
Number of KV heads	5
Embedding dimension	960
Hidden dimension	2560
Total parameters	345M

Table 1: BabyLlama-2 Model Architecture.

Pretraining Approach. The particularity of the BabyLlama-2 model is to be distilled from an ensemble of teacher models, using the distillation loss (1). The teacher models share the same architecture and are pretrained on the same dataset using the standard cross-entropy loss. The student model is then pretrained with the same hyperparameters, using the mean teacher logits $\bar{z}_t$ in the distillation loss $\mathcal{L}_{\text{distill}}(y, z_s, \bar{z}_t)$.

5 Experimental Setup

Dataset. We use the 10 million word BabyLM-2 dataset (Zhuang et al., 2024), that we split into 9.5M train and 0.5M validation splits, as well as the accompanying 10M word “dev” dataset, that we use as a test split. While the validation split is used to perform the hyperparameter optimization,[2]

[2] This choice is dictated by the following logic. A hyperparameter sweep can be viewed as a form of optimization. There-

the test split is used solely for the purpose of reporting the final cross-entropy loss. Each dataset is composed of six files, corresponding each to a different type of (English) language that a child is likely to be exposed to, such as transcribed child-directed speech, children's books, subtitles, or simple Wikipedia. The relative fractions of these files differ slightly between, on the one hand, the train and validation splits and, on the other, the test split, which is therefore slightly out of distribution.

We have experimented with the FineWeb-Edu dataset (Lozhkov et al., 2024) but have observed that models trained on the BabyLM-2 dataset reach better BLiMP scores (see appendix C for more details).

Training. The teacher models are pretrained using the Trainer class from the HuggingFace Transformers library, using the hyperparameters listed in table 2. For the distillation, we use the modified trainer from the original BabyLlama (Timiryasov and Tastet, 2023b), with one, two or three teachers. We use the AdamW optimizer (Loshchilov and Hutter, 2019), with a cosine schedule for the learning rate and 600 warm-up steps. The pretraining hyperparameters have been optimized using a coarse-grained scan, with each parameter being varied independently. The distillation hyperparameters α and T were optimized similarly, while holding the pretraining parameters fixed.

All models share the same Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of 16000 trained on the training split of BabyLM-2 dataset.

Hyperparameter	**Value**
Learning rate	$7 \cdot 10^{-4}$
Number of epochs	8
Batch size	128
Weight decay	5
Distillation T	1
Distillation α	0.5

Table 2: Training and distillation hyperparameters of BabyLlama-2.

Hyperparameter Sweep. To exclude the possibility that the student model BabyLlama-2 outperforms its teachers due to a suboptimal choice of hyperparameters for the teachers, we have performed a comprehensive sweep for the teachers' hyperparameters using the W&B API (Biewald, 2020). We vary the following hyperparameters: the learning rate and its schedule, the Adam parameters (β_1, β_2, ϵ), the batch size, the number of epochs and warm-up steps, the weight decay, the maximum gradient norm, and the attention dropout. We use the Bayesian Optimization and Hyperband (BOHB) (Falkner et al., 2018) parallel sweep algorithm, which stops badly-performing runs early, and we minimize the validation loss at the last epoch. Suitable priors are used for each parameters, usually log-normal or log-uniformly distributed around the values obtained from the coarse-grained scan, with the exception of the attention dropout (uniform) and schedule (discrete). In total, we trained 265 models as part of the sweep, amounting to 26 GPU-days. While the sweep produced some runs that perform noticeable better than the teachers trained with the parameters in table 2, re-training them from a different initial state, but otherwise with the exact same parameters, lead to models that significantly under-performed compared to the initial teachers. Due to this lack of stability with respect to the initialization, we decided to use the original teachers for the distillation procedure.

Benchmarks. We evaluate the performance of the teacher and student models on the benchmarks suggested by the organizers of the BabyLM challenge. Those include zero-shot benchmarks — such as BLiMP (Warstadt et al., 2020), which focuses on linguistic knowledge in English, and EWoK (Ivanova et al., 2024), focusing on world knowledge — as well as the suite of fine-tuning benchmarks SuperGLUE (Wang et al., 2020) about language understanding. For the latter, the fine-tuning hyperparameters are optimized using a separate sweep for each task (totalling 1293 runs and 37 GPU-days). The optimal parameters, listed in table 4, differ significantly from the suggested defaults. See appendix A for further discussion.

Baseline models. The organizers of BabyLM-2 have provided two baseline models: LTG-BERT (Samuel et al., 2023), (encoder-only) and BabyLlama (Timiryasov and Tastet, 2023a) (decoder). Both models were re-trained by the challenge organizers on both the 10M and 100M word datasets. LTG-BERT modifies the original BERT architecture by utilizing the pre-norm variant of the

fore we would consider using the dev split from BabyLM-2 as a violation of the rules of the challenge. Of course, it means that we trained only on 95% of the tokens, and could potentially improve our results further.

transformer with GEGLU feed-forward layers and by disentangling positional information from token embeddings. The highest performing solution of the 2023 edition of the BabyLM challenge, ELC-BERT (Charpentier and Samuel, 2023), is based on this architecture. On the other hand, BabyLlama (the highest-performing decoder model) uses the standard LLaMA architecture (Touvron et al., 2023a), but a modified training procedure, following a similar approach to the one presented here. However, in contrast to BabyLlama-2, it was distilled from two larger teachers with two different architectures (GPT and Llama), and had six times less parameters. Since BabyLlama-2 aims to demonstrate the validity of the ensemble distillation method itself, it uses same-size, homogeneous models in order to remove potential confounding factors. In addition to the baseline models, we vary the number of teachers between 1 and 3, and compare BabyLlama-2 to the ensemble formed by the two teacher models (applying softmax to the averaged logits $\bar{z}_t$ and letting the gradient flow back into both teachers during fine-tuning, with the same training hyperparameters as for BabyLlama-2). When evaluating the original BabyLlama on the SuperGLUE benchmarks, we fine-tune it again using the hyperparameters reported in (Timiryasov and Tastet, 2023a), and successfully reproduce all of its scores.

6 Results

Figure 1 summarizes the performance of the models considered in section 5 with respect to the various evaluation metrics: the cross-entropy loss evaluated on the held-out test set, the BLiMP scores for the "filtered" and "supplement" subsets of evaluation tasks, and the mean SuperGLUE score. The EWoK benchmark is not shown, since the performance of our models and of the baselines trained on 10M words is consistent with random chance, hinting that all these models have extremely limited world knowledge, if any.

Distributions. Violin plots are used in order to quantify the variability across model initializations, with a minimum of 5 runs per model. Each subplot shows a different metric, with the y-axis listing the various models considered: the teacher models, pretrained without distillation; the student models pretrained with one, two or three teachers; the direct ensembles formed by averaging the logits of two teachers; the baseline models for the 2024 BabyLM challenge; and the 265 models from the hyperparameter sweep. No violin is shown for baseline models, since they do not have an associated distribution. Similarly, running fine-tuning benchmarks for all the models from the sweep would have been computationally prohibitive, therefore the SuperGLUE distribution associated with the sweep is not present, with only the best checkpoint being shown.

Models of interest. Instead of, or in addition to the distributions, the performance of various models of interest is plotted using markers. This includes the baseline models, denoted by triangles for BabyLlama and squares for LTG-BERT, with filled markers for baselines pretrained on the 10M word dataset and empty markers for the 100M one. We also indicate with stars the two BabyLlama-2 models that have been submitted to the 2024 edition of the BabyLM challenge. Finally, the cross denotes the best model from the entire sweep, as quantified by its validation loss. The detailed numerical results for the models of interest are listed in table 3, and table 5 further details the SuperGLUE scores of the two submitted BabyLlama-2 checkpoints.

Cross-entropy. The cross-entropy loss is by far the cleanest metric, with a standard deviation across initializations much smaller than the difference between models.[3] It shows a clear and gradual improvement between the teacher models, the student models trained from a single teacher, those trained from two teachers, and those trained from three teachers, although we note that there are diminishing returns as we add more teachers. Even with a single teacher, the improvement is larger that what can be achieved through the hyperparameter sweep. However, looking at the BabyLlama baseline[4], it is clear that this improvement is nowhere near the one resulting from using a ten-fold larger dataset. The cross-entropy loss of the direct ensemble of two teachers is almost as low as for the corresponding model obtained through distillation.

[3] The much larger standard deviation for the sweep comes from including all runs (including early and badly performing runs) instead of just the best runs. The relevant quantity for the sweep is therefore the edge of the distribution. The "best" model is not always located on this edge, since the validation loss does not correlate perfectly with the test loss or the benchmark scores.

[4] The cross-entropy loss is not shown for the LTG-BERT baseline, since it is an encoder-only model trained using masked language modeling, and as such its loss is not comparable to the one discussed here.

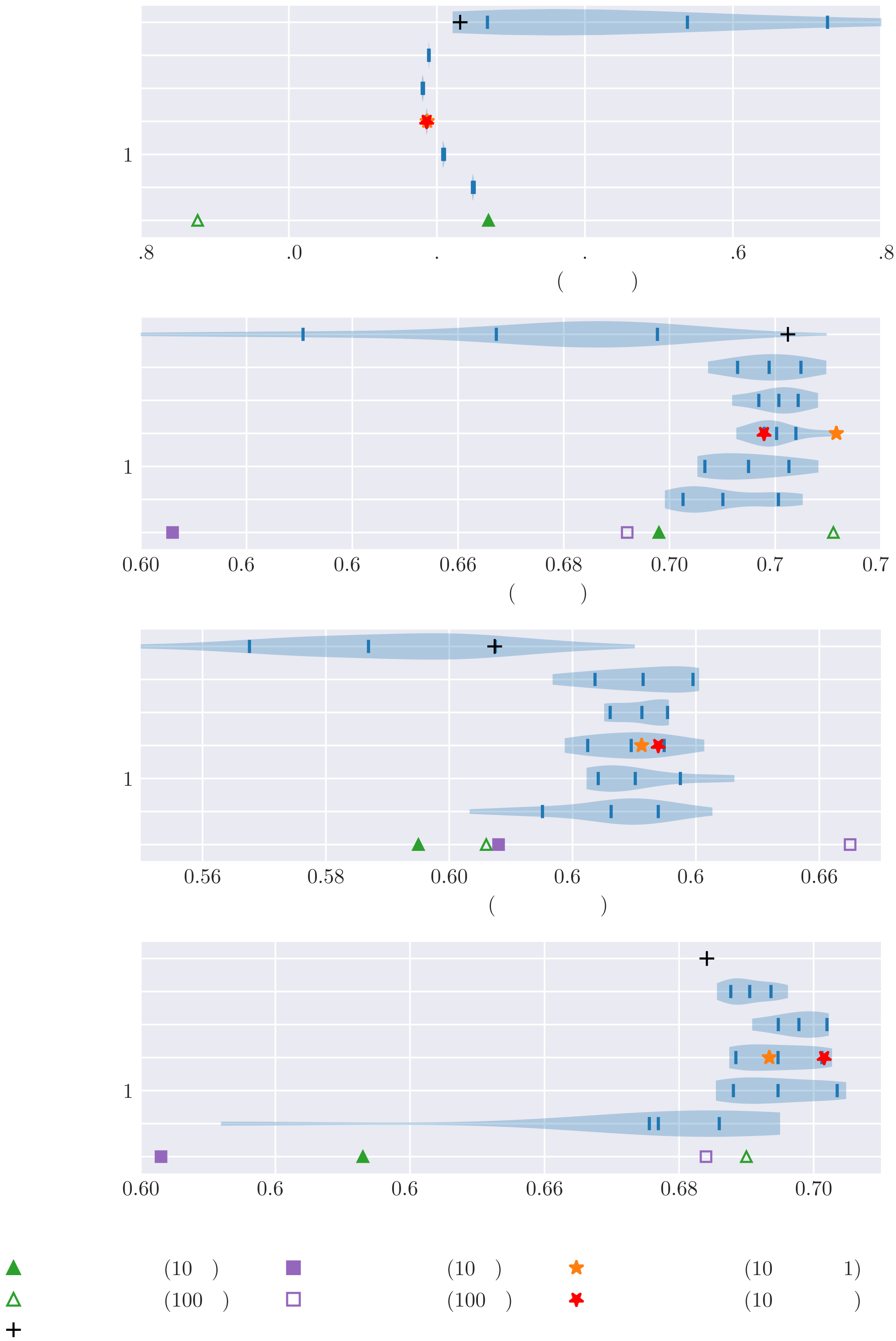

Figure 1: Comparison of the models for each evaluation metric, in the form of violin plots, with ticks denoting the mean and ±1 standard deviation. The baselines are denoted by square and triangle markers, the submitted model (BabyLlama-2) by stars, and the best checkpoint from the entire hyperparameter sweep by a cross. BabyLlama (100M) and LTG-BERT (100M) were trained on the 100M dataset.

Model	BLiMP (filtered)	BLiMP (supplement)	EWoK	SuperGLUE	Macro-average
BabyLlama-2 (run 1)	**73.2**	63.1	50.6	69.3	64.0
Teacher 1	71.9	61.8	50.6	61.2	61.3
Teacher 2	72.1	62.9	50.1	69.5	63.6
BabyLlama-2 (run 2)	71.8	**63.4**	**51.5**	**70.2**	**64.2**
Teacher 1	70.9	62.9	50.4	67.6	62.9
Teacher 2	70.5	62.4	51.1	68.4	63.1
Sweep's best ckpt.	72.2	60.7	50.1	68.4	62.9
BabyLlama (10M)	69.8	59.5	50.7	63.3	60.8
LTG-BERT (10M)	60.6	60.8	48.9	60.3	57.7
BabyLlama (100M)	73.1	60.6	**52.1**	69.0	63.7
LTG-BERT (100M)	69.2	**66.5**	51.9	68.4	64.0

Table 3: Summary of the model scores (in %) across the considered benchmarks. The best scores overall and within the `strict-small` track (10M words maximum) are highlighted.

Benchmarks. The scores on the two BLiMP task sets show a similar trend, but with a significantly higher variability across runs. Because of this, no significant difference is observed between the various distilled or ensemble models. Nonetheless, we can see that the distilled models not only do better than the non-distilled ones, but they tend to achieve this performance more reliably. This is to be contrasted with the performance regression (not shown) that we observed after re-training the best model from the sweep. Direct ensembling leads to similar performance to distillation. Another interesting observation is that despite its much lower cross-entropy loss, the BabyLlama baseline pretrained on 100M words only performs on par with the best BabyLlama-2 model trained on 10M words on the "filtered" subset of tasks, and significantly underperforms on the "supplement" subset. The results are sensibly similar for the SuperGLUE fine-tuning benchmarks, although with much larger variance among the teacher models. Here, again, the distilled models perform more consistently, and they even beat the two baseline models pretrained on the 100M word dataset. Direct ensembling slightly underperforms compared to distillation, but this could be because fine-tuning introduces a dependence on additional hyperparameters, that have not been precisely re-tuned for direct ensembling.[5]

Relation between loss and benchmark performance. The models trained during the hyperparameter sweep allow us to access the relation between the validation loss and BLiMP scores. First, we observe that the loss on our 0.5M word validation set correlates with the loss on the held-out test set with $R^2 = 0.999$. Second, as can be seen from fig. 2, the validation loss explains a significant portion of the variance of the scores: $R^2 = 0.86$ for for BLiMP Filtered and $R^2 = 0.6$ for BLiMP Supplement.

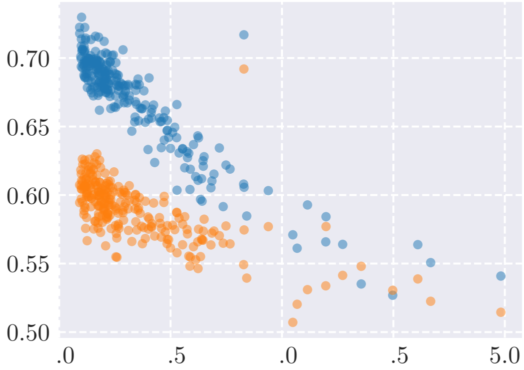

Figure 2: BLiMP scores (averaged over all sub-tasks) as a function of the validation loss. Every circle represents a model from the hyperparameter sweep.

Discussion. The results presented in fig. 1 demonstrate that ensemble distillation from homogeneous teacher models leads to enhanced and more consistent performance across various benchmarks. Notably, BabyLlama-2 often matches or surpasses models pretrained on datasets that are ten times larger. This indicates that the distillation process effectively leverages the knowledge from multiple teachers to compensate for limited data. In addition, the performance of distilled models is consistently as good as, or better than the one of non-distilled models, even when optimizing the hyperparameters of the latter. Therefore, the effect observed in Timiryasov and Tastet (2023a) can-

[5] Naively doubling the fine-tuning learning rate to compensate for the $1/2$ factor resulting from averaging the logits leads to significantly worse performance on SuperGLUE, below that of the teacher models.

not be solely attributed to badly-tuned teacher hyperparameters, and persists even when the student and teachers share the same size and architecture. However, this effect can be difficult to see on the benchmark scores, which are much noisier than the cross-entropy loss. This variability is made particularly evident when looking at the different ordering of the two submitted BabyLlama-2 models across different benchmarks.

Limitations The scalability of the ensemble distillation approach to larger datasets and more substantial model sizes remains unexplored. It is unclear whether the observed benefits will persist or diminish as the scale of data and model parameters increases. Additionally, the exact origin of the improvements from distillation-pretraining remains unclear. Finally, it is not clear whether distillation-pretraining performs significantly better than direct ensembling. Further research, and more sensitive metrics, may be needed to give definitive answers to these question.

7 Conclusions

In this study, we prioritized investigating the robustness of the distillation approach over architectural modifications or dataset curation. Our findings demonstrate that a 345M parameter model, distillation-pretrained on 9.5M words, outperforms models of the same size and architecture pretrained in the usual way. We carried out a systematic analysis to exclude the possibility that the performance gains were due to a single fortunate initialization or suboptimal teacher model hyperparameters. Through an extensive hyperparameter sweep and the training of multiple teacher and student models, we established that distillation-pretraining consistently yields superior performance.

Our results indicate that distillation-pretraining is an effective method for achieving high performance without the need for meticulous hyperparameter tuning, at least within the data-limited regime. The scalability of this approach to larger datasets and model sizes, as well as its applicability to other modalities, remains an open research question.

Acknowledgements

We thank Thea Aarrestad, Erik Dam, Troels C. Petersen, Oleg Ruchayskiy and Raghavendra Selvan for valuable discussions, as well as the anonymous referees for their constructive feedback. This work was supported by a research grant (VIL57416) from VILLUM FONDEN. The work of IT was partially supported by the Carlsberg foundation, and by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 847523 'INTERACTIONS'. Computational resources for this work were partially provided by the SCIENCE AI Centre of Copenhagen University.

A SuperGLUE Fine-tuning

The SuperGLUE suite of benchmarks consists of a number of fine-tuning tasks related to language understanding. Since they involve further model training, the scores crucially depend on the chosen fine-tuning hyperparameters. In table 4, we list the hyperparameters used to fine-tune all our models on the SuperGLUE tasks. These parameters were identified using the BabyLlama-2 checkpoint by performing a separate sweep for each task, and then re-starting the fine-tuning with rounded parameters, in order to check the stability of the found parameters. We have observed that they work well with other model checkpoints, including different versions of BabyLlama-2 and teacher models, suggesting that our hyperparameter selection is robust across different model initializations and pretraining objectives (but not model sizes, since the original BabyLlama had different optimal hyperparameters) and is not overfitted to a specific model or task. The detailed SuperGLUE scores of the two BabyLlama-2 checkpoints submitted to the 2024 BabyLM challenge are reported in table 5.

B Scaling Model Size

We performed initial experiments using a small, 16M version of the model, with the same vocabulary size of 16,000; hidden size 256; intermediate size 1024; 8 layers and 8 attention heads. This model can be fully trained in a few minutes but already achieves decent benchmark scores (without distillation, BLiMP Filtered: 0.68, BLiMP Supplement: 0.58).

To understand the relationship between model size and data requirements, we conducted additional experiments with our 16M and 345M models. We trained these models on random (nested) subsets of the 100M word dataset, ranging from 1M to 100M words each (without re-tuning the hyperparameters). Figure 3 illustrates how the loss decreases as the dataset size increases for both the

Task	Max. learning rate	Batch size	Num. epochs	Weight decay	Schedule	Warm-up steps
CoLA	$1 \cdot 10^{-5}$	32	10	0.15	linear	600
SST-2	$2 \cdot 10^{-6}$	24	2	5	constant	200
MRPC	$1 \cdot 10^{-5}$	1	2	2	cosine	500
QQP	$4.5 \cdot 10^{-6}$	32	6	2	linear	500
MNLI(-mm)	$1 \cdot 10^{-5}$	32	2	1	linear	500
QNLI	$5 \cdot 10^{-6}$	32	2	0.3	cosine	200
RTE	$1 \cdot 10^{-5}$	2	2	10	cosine	200
BoolQ	$2 \cdot 10^{-5}$	8	1	0.1	cosine	200
MultiRC	$1 \cdot 10^{-5}$	8	2	2	cosine	500
WSC	$2 \cdot 10^{-6}$	1	24	0.4	cosine	500

Table 4: List of the hyperparameters selected when fine-tuning BabyLlama-2 on the various SuperGLUE tasks. We do not use early-stopping, since it interfered with BOHB's own early-stopping mechanism. The random seed is 12 for all runs.

Task	Run 1	Run 2
CoLA (MCC)	34.9	31.4
SST-2	85.8	83.5
MRPC (F_1)	82.2	83.8
QQP (F_1)	84.1	84.3
MNLI	74.4	74.3
MNLI-mm	75.3	76.4
QNLI	83.3	83.2
RTE	54.7	61.2
BoolQ	65.9	63.4
MultiRC	64.4	64.9
WSC	57.7	65.4

Table 5: Detailed scores (in %) of the two BabyLlama-2 models on the SuperGLUE tasks. Unless specified otherwise, the listed score is the accuracy. Hyperparameters were optimized for run 1, and then transferred to run 2.

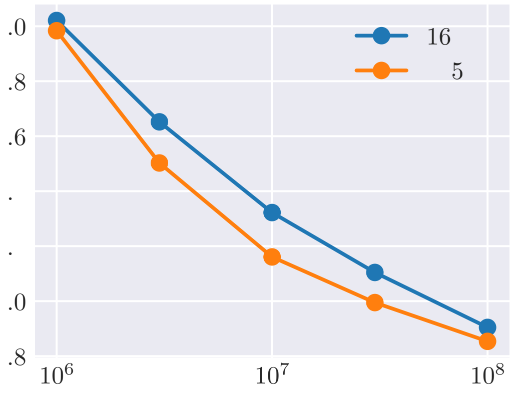

Figure 3: Cross-entropy loss (on the validation split) as a function of dataset size for 16M and 345M models.

16M and 345M models. The 345M model consistently outperforms the 16M model across all dataset sizes, demonstrating that larger models can more efficiently utilize data, hence justifying our choice of the 345M architecture for the final BabyLlama-2 model.

C FineWeb-Edu dataset

Throughout this work, we primarily used the BabyLM-2 dataset. In the early stages, we also experimented with the FineWeb-Edu dataset (Lozhkov et al., 2024), which consists of educational web pages filtered from the FineWeb dataset. We randomly sampled documents containing 20M words (evenly split between the training and validation sets), trained a new tokenizer on this data, and evaluated several variants of the 16M BabyLlama model. The BLiMP scores were consistently lower for models trained on FineWeb-Edu compared to those trained on the BabyLM-2 dataset.[6] We speculate that this lower performance may be due to the limited diversity of examples in FineWeb-Edu, which lacks, for instance, dialogues and non-fiction prose, that are present in BabyLM-2.

References

Loubna Ben Allal, Anton Lozhkov, and Elie Bakouch. 2023. SmolLM - blazingly fast and remarkably powerful. https://huggingface.co/blog/smollm. Accessed: 2023-10-04.

Lukas Biewald. 2020. Experiment tracking with Weights and Biases. Software available from wandb.com.

[6] We do not report specific numbers here since the method for averaging the scores has changed since these experiments were conducted.

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

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Stefan Falkner, Aaron Klein, and Frank Hutter. 2018. BOHB: robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. 2018. Born again neural networks. In *International conference on machine learning*, pages 1607–1616. PMLR.

Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. 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.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv e-prints*, arXiv:1503.02531.

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.

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. *Preprint*, arXiv:1907.11692.

Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. 2024. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Forty-first International Conference on Machine Learning*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. 2024. FineWeb-Edu.

Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, et al. 2023. FinGPT: Large generative models for a small language. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2710–2726.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. *arXiv e-prints*, arXiv:2305.16264.

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

Inar Timiryasov. 2023. Speed of Llama. `https://timinar.github.io/posts/speed-of-llama/`.

Inar Timiryasov and Jean-Loup Tastet. 2023a. Baby Llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 279–289.

Inar Timiryasov and Jean-Loup Tastet. 2023b. BabyLlama GitHub repository.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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 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.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Chengxu Zhuang, Ethan G Wilcox, Alex Warstadt, and Aaron Mueller. 2024. BabyLM_2024.