Big Impact from Small Models: The Power of Curated Corpora

Anonymous ACL submission

Abstract

In this paper, we present an alternative approach to language model training that emphasizes data quality over sheer volume. Using the authoritative *Encyclopaedia Britannica* as corpus, we first limit the size to a 10-million-word dataset. We then expand the training of a specialized model for encyclopedic content generation with the complete 38-million-word corpus. Central to our approach is the use of knowledge distillation, which allowed us to train compact student models guided by larger teacher models, achieving high performance while significantly reducing model complexity. Building on the BabyLlama architecture (Timiryasov and Tastet, 2023), our findings reveal that high-quality, curated data combined with effective distillation techniques can facilitate efficient and effective learning. This work highlights promising directions for resource-constrained applications and specialized domain modeling. We will release our programs and models if this paper is accepted.

1 Introduction

The development of large language models (LLMs) has been predominantly driven by massive datasets, often exceeding a trillion tokens. This approach follows the Chinchilla scaling law, which suggests an optimal ratio of 20 tokens per model parameter (Hoffmann et al., 2022). However, this scaling trajectory presents several critical challenges for the field. First, the availability of high-quality training data has become a significant bottleneck. As models grow larger, finding sufficient high-quality data becomes increasingly difficult, often forcing researchers to rely on noisier sources. Second, current approaches demonstrate remarkably low sample efficiency compared to human language acquisition.

These challenges point to a fundamental question: Can we develop more efficient training approaches by prioritizing data quality over quantity? In this paper, we show that carefully curated, expert-reviewed content might enable more efficient learning than the vast but heterogeneous datasets currently in use. In our experiments, we utilize the *Encyclopaedia Britannica* as our training corpus, pursuing two distinct but complementary research directions.

In the first experiments, we followed the BabyLM Challenge setup (Warstadt et al., 2023), which constrains models to training on just 10 million words. This constraint aligns with our interest in efficient learning from limited but high-quality data. Our second set of experiments explores specialized model development using the complete *Encyclopaedia Britannica* corpus of 38 million words, aiming at creating a model specifically tuned for encyclopedic content generation.

2 Related Work

The introduction of the Transformer architecture (Vaswani et al., 2017) and its pretraining with BERT (Devlin et al., 2019) revolutionized the NLP field. However, as they improved in performance, they also increased in size and in training corpus needs. BERT used 3.3B words crawled from BookCorpus and English Wikipedia. The issue became bigger with newer models: RoBERTa (Liu et al., 2019) was trained on more than 30B words, XLNet (Yang et al., 2020) on 33B words, T5 (Raffel et al., 2023) and OPT (Zhang et al., 2022) on 170B and 180B words respectively, GPT-3 (Brown et al., 2020) on 300B words, while Chinchilla (Hoffmann et al., 2022) and Llama (Touvron et al., 2023) on approximately 1.4 trillion words. Another issue is that while the datasets used have exploded in size, they are usually not available in public and do not aid reproducibility.

Previous works showed that high quality data can improve model performance, even if they are available in lower quantity. In Taylor et al. (2022), the authors demonstrated that research papers, encyclopedias, lecture notes, equations, and chemistry compounds give a significant boost in performance and the model's knowledge. While the quality of the dataset helped, the model was still trained on 106 billion tokens. Later, Gunasekar et al. (2023) trained a model that outperforms famous state-of-the-art models using only 7B tokens derived from textbooks. In this work, we explore how a high quality corpus of even smaller size, like an encyclopedia, can further diminish training needs, while preserving performance.

3 Dataset

Source Selection and Rationale. Our choice of the *Encyclopaedia Britannica* as a training corpus was driven by several key considerations. First, its content undergoes rigorous expert review, ensuring high accuracy and consistent quality. Second, it maintains a formal and standardized writing style on diverse topics, providing excellent examples of structured knowledge presentation. Third, its comprehensive coverage of human knowledge makes it suitable for both general language learning and specialized content generation.

This work uses the 10th edition from 1911, comprising 29 volumes. Although there are more recent editions, this particular version offers distinct advantages for our work. It is readily accessible in digital format, maintains consistent editorial standards throughout its volumes, and its historical nature provides a well-defined temporal boundary for knowledge scope. In addition, its public domain status facilitates unrestricted research use and reproducibility of our findings.

Data Processing Pipeline. Our processing pipeline consists of a sequence of stages to ensure data quality and consistency. We scraped the individual articles from Wikisource, converting them from their original format into structured JSON files. This initial conversion preserves article metadata – including titles, section headers, and publication dates – while extracting the main content for further processing.

The next stage involves text cleaning and standardization through a series of regular expressions. Our cleaning process systematically removes special characters and formatting artifacts from the text while standardizing all punctuation and spacing patterns. We eliminate references and citations to maintain focus on the primary content.

To facilitate model training, we incorporate explicit article boundary tokens (`<s>` and `</s>`) between entries. These boundary markers enable the model to learn natural document boundaries, maintain coherent article generation, and prevent content blending between adjacent articles. These tokens also provide useful attention anchors during the generation phase, helping the model structure its output.

Dataset Organization. We created two different experimental setups. We selected a 12-million-word random subset of the corpus, ensuring balanced topic coverage and maintaining the encyclopedic style. We divided it into training (10 million words) and evaluation (2 million words) sets. For our encyclopedia generation track, we used the entire corpus of approximately 38 million words split into training (37 million words) and validation (1 million words) sets. All splits are created at article boundaries.

Tokenization. We applied a Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of 16,000 tokens, trained exclusively on our training set. This vocabulary size aligns with recent research suggesting that moderate vocabulary sizes can be optimal for specialized domains. The tokenizer is trained separately for each setup.

The tokenized text is segmented into sequences of 128 tokens. This means that longer articles may span multiple sequences, while shorter articles might be combined within a single sequence. For sequences shorter than 128 tokens, we apply padding with special `<pad>` tokens to maintain uniform input dimensions. The article boundary tokens ensure that semantic coherence is preserved across these mechanical divisions, helping the model distinguish between true article boundaries and arbitrary segmentation points.

4 Model Architecture and Training

4.1 Knowledge Distillation Framework

Our training methodology applies a knowledge distillation framework (Bucila et al., 2006; Hinton et al., 2015), where an ensemble of teacher models guides the training of more compact student models. This approach is particularly well-suited to our objective of extracting maximum value from a limited, high-quality dataset. Building upon the foundational work of the BabyLlama authors (Timiryasov and Tastet, 2023), we aim at replicating their method while extending it through additional experiments to further explore its potential.

Teacher Models. We utilize two teacher models with complementary architectures to achieve robust knowledge distillation.

The first teacher is a GPT-2 model with 24 layers and 16 attention heads, featuring an embedding dimension of 1536 and an intermediate dimension of 6144. With approximately 705 million parameters, this model offers substantial capacity to capture intricate patterns. It was trained for six epochs using a batch size of 256 and a maximum learning rate of 2.5×10^{-4}. A cosine learning rate scheduler with 300 warm-up steps was employed to ensure smooth optimization.

The second teacher is a Llama model comprising 24 layers and 8 attention heads. It has a hidden dimension of 1024 and an intermediate dimension of 3072, totaling roughly 360 million parameters. This model strikes a balance between computational efficiency and representational power. Training was conducted for four epochs with a batch size of 128 and a maximum learning rate of 3×10^{-4}, again utilizing a cosine learning rate scheduler with 300 warm-up steps.

We pretrained both models exclusively on our encyclopedia corpus to align their knowledge with the target domain. During training, we computed the validation loss after each epoch using a smaller random subset of the evaluation set (8192 samples).

Distillation Process. The distillation process leverages both hard targets from the training data and soft targets generated by the teacher models. The soft targets, which are probability distributions over the output classes, encapsulate relational information between classes learned by the teacher models. To enhance this knowledge transfer, we divided the logits by a temperature value T with $T = 2$.

The student models are trained to optimize a combined loss function that aligns their outputs with both the hard targets (ground truth labels) and the soft targets (teacher predictions). The loss function is defined as: $L = \alpha L_{\text{CE}} + (1 - \alpha) L_{\text{KL}}$. Here, L_{CE} is the cross-entropy loss between the student's predictions and the true labels, while L_{KL} represents the Kullback-Leibler (KL) divergence between the student's and teacher's softened probability distributions.

We set $\alpha = 0.5$ to balance the two objectives. By combining these two losses, our approach encourages the student models to benefit from the complementary strengths of the teacher models while retaining alignment with the original task objectives. This dual alignment is critical for producing an effective student model.

4.2 Student Models

In both the BabyLM Challenge and the encyclopedia generation tasks, we employed a unified student model architecture designed to balance the model capacity with the constraints of available training data. We trained the architecture from scratch using a distillation loss. Both tasks used identical model configurations and training strategies to maintain consistency across the domains.

The student model is based on a Llama architecture with 16 transformer layers, each containing 8 attention heads. It features a hidden dimension of 512 and an intermediate dimension of 1024, resulting in approximately 58 million parameters. We trained it with a batch size of 32 and a maximum learning rate of 2.5×10^{-4}. We employed a cosine learning rate schedule with 200 warm-up steps. We applied regularization techniques, including dropout and weight decay, to prevent overfitting. Additionally, we used gradient clipping to maintain stable training dynamics. In total, the training process took about 12 hours on one A-100 Nvidia GPU.

5 Experimental Results

To evaluate our models, we primarily used the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020), a zero-shot evaluation suite designed to test a language model's ability to handle linguistic phenomena such as syntax, semantics, and morphosyntax. BLiMP contains 67,000 minimal sentence pairs across 67 tasks. Models are scored based on their ability to assign higher probabilities to grammatically correct sentences than to ungrammatical ones. Additionally, we used the BLiMP Supplemental benchmark, an extension with more diverse and challenging tasks.

For the evaluation, we used the BabyLM pipeline, which outputs the average scores across the tasks for each benchmark. We compared our BabyLM Challenge student model, trained on 10 million words, against baseline models provided by the Challenge organizers: OPT (125M parameters), RoBERTa (125M parameters), and T5 (222M parameters). These baselines were trained on a cu-

Model	BLiMP	BLiMP Suppl.
OPT$_{125M}$	62.6	**54.7**
RoBERTa$_{125M}$	**69.5**	47.5
T5$_{222M}$	58.8	43.9
GPT-2$_{705M}$	64.5	49.3
LLama$_{360M}$	62.8	48.2
student$_{58M}$	64.7	47.6
student$_{58M+GPT-Jteacher}$	64.6	47.1
studentFULL$_{58M+38Mdataset}$	66.4	50.3

Table 1: Results on the BLiMP benchmarks.

Model	BLiMP	BLiMP Suppl.
OPT$_{125M}$	0.50	0.43
RoBERTa$_{125M}$	0.55	0.38
T5$_{222M}$	0.26	0.19
GPT-2$_{705M}$	0.09	0.06
LLama$_{360M}$	0.17	0.13
student$_{58M}$	**1.11**	**0.82**
student$_{58M+GPT-Jteacher}$	1.11	0.81
studentFULL$_{58M+38Mdataset}$	**1.14**	**0.86**

Table 2: Scores scaled by the number of parameters.

rated 10-million-word dataset composed of child-directed speech, storybooks, simple Wikipedia, and transcribed speech.

We also compared the results of two teacher configurations: GPT2/Llama with GPT-J/Llama. The difference between GPT-J and GPT-2 is that GPT-J employs rotary positional embeddings, while GPT-2 uses trainable positional embeddings. We trained the student model on the 10-million word *Encyclopaedia* dataset. Table 1 shows the raw results, while Table 2 shows the relative performance of each model scaled by the number of its parameters.

Our BabyLM Challenge student model achieves a performance comparable to those of the baselines, outperforming OPT and T5 on the BLiMP benchmark and T5 on the Supplemental benchmark, while slightly trailing RoBERTa and OPT. Notably, despite having less than half the parameters of these baselines, the student model achieves competitive results, highlighting its efficiency.

When compared to the teacher models, the student outperforms both GPT-2 and Llama on BLiMP, although it slightly underperforms on the Supplemental benchmark. Considering that the teachers have 12x and 6x more parameters than the student, this underscores the effectiveness of knowledge distillation in transferring knowledge efficiently.

The student model trained with GPT-J slightly underperformed compared to the GPT-2 teacher. This suggests that GPT-2's trainable positional embeddings provide a distinct advantage during distillation. Lastly, the student model trained on the full 38-million word Encyclopedia dataset shows a 2–3% improvement on both benchmarks, demonstrating the benefit of additional high-quality training data.

When we observe Table 2 with the relative performance, it is evident that our model significantly outperforms the other models, having more than double the score of the second ranked model.

We finally give two examples of generation:

```
Prompt: "<s> Italy is"
Generated:  Italy is divided into two
    great parts, the north and the south
    , and the central and south-west;
    while the south, including the
    greater part of the peninsula, and
    the south, the eastern and the
    northern parts, are separated by the
     great plain of Italy. The whole of
    the western and western part of the
    peninsula, except the peninsula of
    the Mediterranean, is a very fertile
     plain, and in some places has been
    made the capital of the kingdom of
    Italy. The whole of northern Italy
    is traversed by the main line of the
     Italian railway.

Prompt: "<s> London was"
Generated: London was the capital of the
     district of the East Anglian
    kingdom of North Wales, the chief
    town of which was Beverley. It was
    the seat of the royal family of the
    earls of South Wales, who held it in
     the time of Edward I. It was a
    royal burgh of barony, and held the
    manor of Fordun in the time of
    Edward I.
```

While the model hallucinates sometimes, in both cases, it manages to capture the context of the prompt (country and town) and generate content. Further approaches would finetune the model for question-answering or use a RAG system to produce more accurate results.

6 Conclusion

Our two experimental setups highlight the potential of training highly effective language models on smaller, high-quality datasets. Knowledge distillation significantly reduced the number of model parameters while maintaining performance. This generation model shows that curated data and teacher-student training frameworks can offset limited dataset sizes.

Limitations

Large language models may generate factual errors or show bias. We observed them in the form of hallucinations. This bias may also come from the corpus we used for training, the *Encyclopaedia Britannica*, that includes old fashioned viewpoints and outdated information. Many scientific theories have also evolved since it was published.

Our evaluation is also limited and a comprehensive study would include more benchmarks and more models.

Ethics Statement

We identified a few potential risks including:

1. The LLMs may generate misleading content or hallucinate;

2. The dataset we used for training includes biases and outdated information that in some cases may generate disrespectful or disparaging content;

Acknowledgments

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Knowledge Discovery and Data Mining*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *Preprint*, arXiv:2211.09085.

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

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

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

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

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2020. Xlnet: Generalized autoregressive pretraining for language understanding. *Preprint*, arXiv:1906.08237.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.