

Stable LM 2 1.6B Technical Report

Marco Bellagente* Jonathan Tow* Dakota Mahan Duy Phung Maksym Zhuravinskyi
 Reshinth Adithyan James Baicoianu Ben Brooks Nathan Cooper Ashish Datta
 Meng Lee Emad Mostaque Michael Pieler Nikhil Pinnaparaju Paulo Rocha
 Harry Saini Hannah Teufel Niccolo Zanichelli Carlos Riquelme

Stability AI Language Team*

Abstract

We introduce StableLM 2 1.6B, the first in a new generation of our language model series. In this technical report, we present in detail the data and training procedure leading to the base and instruction-tuned versions of StableLM 2 1.6B. The weights for both models are available via Hugging Face for anyone to download and use ^{†‡}. The report contains thorough evaluations of these models, including zero- and few-shot benchmarks, multilingual benchmarks, and the MT benchmark focusing on multi-turn dialogues. At the time of publishing this report, StableLM 2 1.6B was the state-of-the-art open model under 2B parameters by a significant margin. Given its appealing small size, we also provide throughput measurements on a number of edge devices. In addition, we open source several quantized checkpoints and provide their performance metrics compared to the original model.

1 Introduction

Following the development of the Transformer architecture [71], a remarkable number of proprietary and open-source large language models have been trained and deployed. While countless new ideas and artifacts are announced weekly or daily, some key aspects remain opaque – especially around the most powerful models. Often, the training data is not disclosed. This poses a fundamental challenge in times where society demands transparency as it faces a brand-new disruptive technology that is hard to inspect and audit. In this report, we explain in a reproducible manner how to train a modest-size but state-of-the-art language model. All the data we used is public (see Table 1) and its training required around 92k GPU hours of training – worth around \$322k on popular cloud providers (assuming \$3.5 per GPU hour). We hope this work contributes to the open AI community and helps set the standard for upcoming transparent models.

This report is organized as follows: Section 2 details the process of pre-training Stable LM 2 1.6B. We devote Section 3 to fine-tuning and human preference alignment. Section 4 presents model evaluations on standard downstream benchmarks. Compiling and running inference Stable LM 2 on several edge devices is outlined in Section 5. We consider several follow-up directions in Section 6. Carbon emissions and societal impacts related to the training and release of Stable LM 2 are covered in Section 7. Finally, Section 8 concludes and summarizes this work.

2 Model Pre-Training

The first stage in training large language models (LLMs) focuses on learning to predict the next token in a sequence using a vast and diverse array of data sources. We refer to this stage as *pre-*

*Equal contribution. Correspondance to: {marco.bellagente, jonathantow, carlos.riquelme}@stability.ai

[†]https://huggingface.co/stabilityai/stablelm-2-1_6b

[‡]https://huggingface.co/stabilityai/stablelm-2-zephyr-1_6b

training. It enables models to build general-purpose internal representations suitable for basic language capabilities and even more advanced generation and comprehension tasks. In fact, it has been hypothesized that the majority of model knowledge and capabilities are learned during pre-training [88]. In this section, we introduce the design principles and ablations that influenced the creation of our training set, as well as details about the model architecture and training procedure. While many similar reports exist for other cutting-edge models, they often omit critical details, such as the particular data sources, sampling weights, or the complete set of ablations they performed. As a result, the open-source community cannot accurately reproduce these models. On the other hand, we present a fully transparent log of our model training details. We are confident that researchers and practitioners will find valuable insights in this comprehensive account.

Dataset	Sampling Weight	Num Tokens	Epochs	Category
Arxiv	0.0079	15,852,446,656	0.75	Academic
PubMed	0.012	24,126,600,626	1.0	Academic
S2ORC	0.0318	63,910,567,858	1.0	Academic
PhilPapers	0.0013	2,591,115,796	4.0	Academic
BookCorpusOpen	0.0045	9,135,954,810	6.0	Books
PG-19	0.0091	18,293,226,468	4.0	Books
FanFics	0.0018	3,644,583,700	4.0	Books
Cultura-X (EN)	0.2521	506,625,211,049	0.72	Web
Cultura-X (ES)	0.0155	31,241,701,156	0.4	Web
Cultura-X (DE)	0.0152	30,628,813,934	0.32	Web
Cultura-X (FR)	0.0097	19,466,611,808	0.26	Web
Cultura-X (IT)	0.0096	19,202,903,491	0.4	Web
Cultura-X (NL)	0.0097	19,511,322,386	0.62	Web
Cultura-X (PT)	0.01	20,063,968,694	0.78	Web
C4	0.0855	171,782,178,108	1.0	Web
OpenWebText2	0.0130	26,161,864,434	3.0	Web
RefinedWeb	0.3292	661,591,178,339	1.15	Web
StackExchange	0.0231	46,302,993,820	2.5	Social
HackerNews	0.0019	3,817,060,582	2.0	Social
EuroParl	0.0023	4,678,506,882	3.0	Law
FreeLaw	0.0088	17,694,697,577	1.2	Law
PileOfLaw	0.0063	12,663,061,642	0.75	Law
DM Math	0.0066	13,321,872,138	3.0	Math
AMPS	0.0011	2,126,719,278	6.0	Math
OpenWebMath	0.0132	26,530,341,292	2.0	Math
RedPajama Wiki	0.0363	72,866,870,472	3.0	Wiki
Starcoder	0.0724	145,586,775,301	0.74	Code
Restruct-v1	0.0102	20,412,655,632	3.0	Instruction
Total	1	2,009,831,803,933	-	

Table 1: The complete Stable LM 2 training set with sampling weights. The tokens count refers to the **Arcade100k** tokenizer introduced in Sec. 2.3. The number of tokens in the table already includes the number of epochs shown next to it. For instance, in the case of BookCorpusOpen, we use around 9B tokens, corresponding to 6 epochs of the original dataset (that is, each epoch is around 1.5B tokens). Similarly, if epochs are below one, the number of tokens shown is a subset of the total dataset.

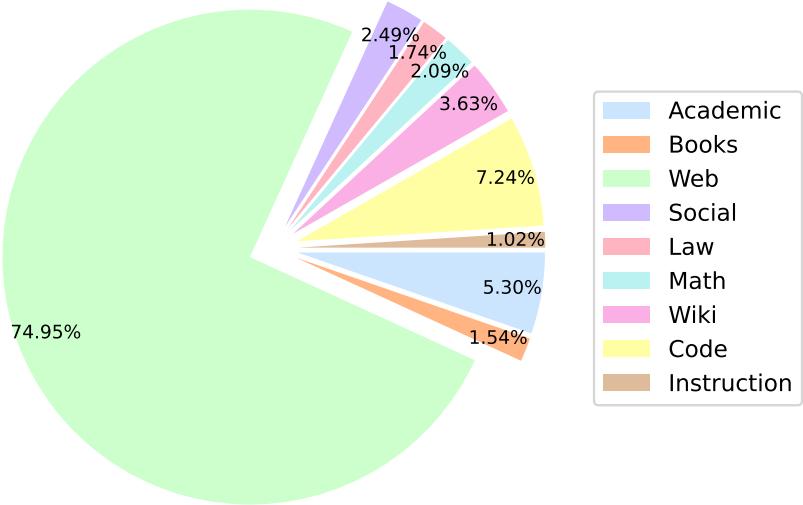


Figure 1: Percentage of effective training tokens by domain in the Stable LM 2 pre-training dataset.

2.1 Training

We train Stable LM 2 to predict the next token following standard autoregressive sequence modeling [58]. We train our model from scratch with a context length of 4096 and benefit from the efficient sequence-wise parallelism optimizations of FlashAttention-2 [17, 16]. Training is performed in BFloat16 mixed precision while keeping all-reduce operations in FP32. [10, 74] found it beneficial to add a z-loss regularization term on the softmax normalizer, $z_{\text{loss}} \propto \log^2 Z$, to mitigate instability stemming from output logits divergence. While it did not hurt performance in our ablations, the improvements to stability were minimal. Accordingly, it was not applied in the final run. We employ a standard AdamW optimizer with the following hyperparameters: $\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 1e - 8, \lambda(\text{weight decay}) = 0.1$. Sec. 2.5 offers details regarding the custom learning rate scheduler that we applied.

2.2 Data

Model performance is affected by pre-training data design decisions, including both the source selection and the sampling weights [48]. Our approach is close to that of [67]: the majority of our training data consists of sources utilized in the training of other LLMs, such as RefinedWeb [57], subsets of the Pile [22], RedPajama [13] and the Stack [39]. We supplement these with OpenWebText [24], OpenWebMath [56], and parts of CulturaX [54]. While inspecting randomly sampled documents from the mC4 subset of CulturaX, we encountered frequent HTML boilerplate and decided to remove this portion entirely, finally keeping only the OSCAR subset. Additionally, we incorporate FanFics[§], a subset of 50k documents from *fanfiction.net* selected by lowest perplexity scores according to a KenLM[¶]. Finally, following [81], we restructure several raw datasets into rich fixed forms for downstream tasks such as summarization, question-answering, sentiment analysis, etc., and add instruction data from [47], the aggregate of which we collectively call Restruct-v1. The list of sources used in Restruct-v1 is made available in Tab. 11. Stable LM’s training set is comprised entirely of open-source datasets compatible with commercial usage, most of which are hosted on the Hugging Face Hub. The only exception to the latter aspect (HFH), Restruct-v1, can easily be reproduced by following the approaches and prompt templates provided by [81].

[§]<https://huggingface.co/datasets/atom-in-the-universe/fanfics-10k-50k>

[¶]<https://huggingface.co/edugp/kenlm>

Parameters	Hidden Size	Layers	Heads	Sequence Length
1,644,417,024	2048	24	32	4096

Table 2: Stable LM 2 model architecture.

Data Parallel Degree	Micro Batch Size	Gradient Accumulation Steps	Activation Checkpointing
512	2	2	disabled

Table 3: Stable LM 2 training configuration.

Carefully selecting the mixture proportions of the various data domains is critical, particularly with respect to the amount of non-English and code data. We trained several models on different mixes and evaluated them on downstream benchmarks to pick our final dataset. The full set of ablations is available in Appendix A, together with rationales for selecting these particular mixes. Based on the results of the ablations, we trained our model with the mix shown in Table 1, which accounts for around 2 trillion tokens. Note that it includes multilingual data in German (DE), Spanish (ES), French (FR), Italian (IT), Dutch (NL), and Portuguese (PT). The split of our dataset across different domains is visualized in Fig. 1.

2.3 Tokenizer

We use **Arcade100k**, a BPE tokenizer extended from OpenAI’s `tiktoken.cl100k_base` to include special tokens for code [40] and digit-split handling [45, 4]. The vocabulary consists of 100,289 tokens and is padded to the nearest multiple of 64 (100,352) during training to meet the recommended Tensor Cores alignment on NVIDIA A100 devices. In preliminary experiments, we did not observe statistically significant deviations in downstream natural language performance tasks when compared against a hyperparameter matching model trained with the smaller GPT-NeoX tokenizer [6]. Increased compression rates for code and non-English languages influenced our decision to use **Arcade100k** over existing tokenizers.

2.4 Architecture and Training Layout

The model is a causal decoder-only transformer similar in design to the LLaMA architecture [67]. Table 2 shows some of the key architectural details. In particular, the main differences with respect to LLaMA are the following:

- **Position Embeddings.** Rotary Position Embeddings [66] applied to the first 25% of head embedding dimensions for improved throughput following [6].
- **Normalization.** LayerNorm [3] with learned bias terms as opposed to RMSNorm [84].
- **Biases.** We remove all bias terms from the feed-forward networks and multi-head self-attention layers, except for the biases of the key, query, and value projections [4].

Stable LM 2 1.6B is trained on 64 Amazon P4d instances comprising 512 NVIDIA A100 (40GB HBM2) GPUs. The size of our model, together with ZERO stage 1 distributed optimization [61], eliminates the need for model sharding. Still, different triplets of **micro batch size**, **gradient accumulation steps**, and **activation checkpointing** granularity lead to different speed metrics. Following the recommendations in [28], we obtain our final configuration by finding the micro-batch size that allows us to completely remove activation checkpointing. We then determine the gradient accumulation steps based on our target batch size and **data parallel** degree. We employ a batch size of 8,388,608 tokens, based on the observations in Appendix D. With the setup in Table 3, we achieve \approx 170 TFLOPs/s per device, or 54.5% model flops utilization (MFU). A higher hardware utilization of \approx 200 TFLOPs/s (64% MFU) can be trivially achieved by decreasing the degree of data parallelism and correspondingly increasing the number of gradient accumulation steps, at the cost of an increased iteration time.

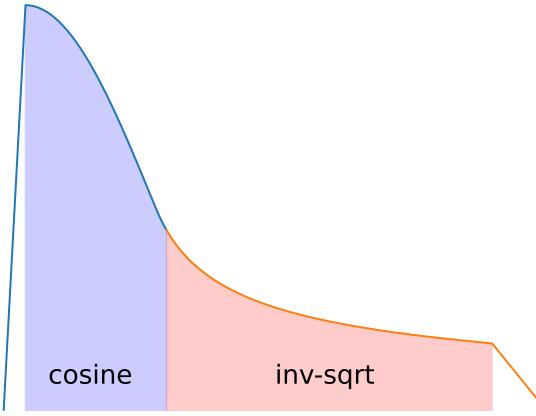


Figure 2: Multi-stage infinite scheduler proposed and applied in this work.

2.5 Learning Rate Scheduler

We propose a new learning rate scheduler. It consists of multiple stages and is designed to favor flexibility in terms of continued pre-training. We begin by linearly increasing the learning rate to its max value of $1e - 3$ over 9720 steps. This *warm up* stage is followed by the main training phase in which the learning rate is decreased according to Eq. 1:

$$\begin{cases} m + \frac{(M - m)}{2} * \left[\cos \left(2\pi * \frac{i}{N} \right) + 1 \right] & \text{if } i \leq N/4 \quad \text{cosine decay} \\ \frac{\alpha}{\sqrt{i + \beta}} & \text{if } i > N/4 \quad \text{rsqrt decay} \end{cases} \quad (1)$$

where m and M are respectively the minimum and maximum learning rates, i is the current step, and N is the total number of steps. The free parameters α and β have been arbitrarily chosen to enforce the continuity of the scheduler and its derivative at $i = N/4$. We finalize training by linearly cooling the learning rate down to zero over 80k steps, which corresponds to around 670B tokens. The full scheduler is illustrated in Fig. 2; further details and ablations can be found in Appendix B.

3 Fine-tuning and Alignment

Following pre-training, we further develop the conversational skills of our model via a fine-tuning stage that consists of three main steps: supervised fine-tuning (SFT), direct preference optimization (DPO), and self-knowledge learning. Importantly, we do not use multilingual data at this stage. We now describe each of them in detail, and in Section 4 we report the results after all three steps.

3.1 SFT

The first step is supervised fine-tuning. We fine-tune the pre-trained model on a number of instruction datasets publicly available on the Hugging Face Hub. In particular, we use the following *conversational* datasets: UltraChat [18], WizardLM [77], SlimOrca [41], ShareGPT [72], Capybara [15], Deita [46], and MetaMathQA [80]. We removed any samples that exceeded eight turns, leading to a total of 826,938 samples.

We train our SFT models for three epochs using a cosine learning rate scheduler. A warm-up phase of 10% of the training duration is employed before reaching the peak learning rate of $5e - 5$. We set the global batch size to 512 sequences and pack inputs into sequences of up to 4096 tokens in length.

3.2 DPO

Direct Preference Optimization (DPO) [58] has been a fundamental tool in recent strong models such as Zephyr-7B [70], Neural-Chat-7B, and Tulu-2-DPO-70B [32]. Accordingly, after applying SFT, we

Model	Size	Avg	ARC	HS	MMLU	TQA	Wino	GSM
phi-2†	2.8B	61.3	61.1	75.1	58.1	44.5	74.4	54.8
stablelm-2-zephyr-1_6b†	1.6B	49.7	43.3	69.3	41.8	45.6	63.6	34.8
phi-1_5†	1.3B	47.7	52.9	63.8	43.9	40.9	72.2	12.4
stablelm-3b-4e1t	2.7B	46.6	46.6	75.9	45.2	37.2	71.2	3.3
Qwen-1.5-1.8B	1.8B	46.6	37.9	61.4	46.7	39.4	60.3	33.6
gemma-2b	2.5B	46.5	48.5	71.7	41.7	33.1	66.8	17.4
stablelm-2-1_6b	1.6B	45.3	43.3	70.5	38.9	36.8	64.6	17.4
gemma-2b-it†	2.5B	42.7	43.9	62.7	37.6	45.8	60.9	5.5
open_llama_3b	3B	40.3	39.9	71.6	27.1	34.8	67.0	0.9
falcon-rw-1b	1.3B	37.1	35.1	63.6	25.3	35.9	62.0	0.5
TinyLLama-1.1B	1.1B	36.4	33.9	60.3	26.0	37.3	59.5	1.4

Table 4: Comparison of Open LLM leaderboard evals. † denotes aligned models. Note that in this table, as well as in Tab. 5, we marked the Phi series of models [26] as *aligned*. While we acknowledge that they may have only been pre-trained, by the nature of the training data used, and the self-disclaimed intended use for QA and chat formats, we believe this to be a fair statement.

Model	All	EN	DE	ES	FR	IT	NL	PT
stablelm-3b-4e1t	41.7	50.9	39.7	40.2	41.2	41.1	36.7	41.9
stablelm-2-zephyr-1_6b†	41.5	49.5	40.2	40.0	39.8	39.9	38.8	42.0
stablelm-2-1_6b	40.5	48.7	39.1	39.0	39.3	38.8	37.8	41.2
gemma-2b	39.8	48.6	38.3	38.7	38.7	38.4	35.1	40.5
gemma-2b-it†	38.2	49.4	36.8	38.0	37.5	35.5	32.1	38.1
open_llama_3b	37.5	47.3	35.2	36.4	37.6	37.1	32.2	36.8
Qwen-1.5-1.8B-Chat	35.5	46.2	33.2	35.1	34.3	33.2	30.9	35.7
Qwen-1.5-1.8B	34.8	46.3	31.8	34.0	34.2	32.8	29.7	35.0
TinyLLama-1.1B	34.8	42.4	33.0	33.8	34.7	33.5	31.0	35.0
phi-2†	34.6	55.8	29.0	34.3	32.9	29.9	27.1	33.4
falcon-rw-1b	29.9	42.2	27.4	28.6	28.3	28.0	25.9	29.1
phi-1_5†	29.7	47.1	25.3	28.7	26.8	27.2	24.1	29.0

Table 5: Multilingual evaluations. As in the previous table, † denotes aligned models.

aligned the resulting model via DPO. We use two datasets at this stage: UltraFeedback [14] and Intel Orca Pairs. We filter the datasets by removing pairs with tied ranking, pairs with duplicated content, and pairs in which the score for the chosen response is less than eight out of ten. We train the model with DPO following the Zephyr recipe [70] and borrowing most of its hyperparameters, except for β , which we lower to 0.01, and the learning rate, which we lower to $4e - 6$, both of which helped improve the stability of training and final performance. This stage of training is performed using the Alignment Handbook [69].

3.3 Self-Knowledge

The model after the output of the DPO [58] stage does not have knowledge about who created it, or even what limitations a language model has. To remedy this, we were inspired by the data generation method of Reinforcement Learning from Contrast Distillation (RLCD) [78] and the training method of Conditioned Reinforcement Learning Fine-Tuning (C-RLFT) [72], which we apply to self-knowledge training.

To generate the initial prompts, we use the base model to generate 10k random first messages to a language model with no duplicates. To generate continuations, we use a few-shot prompt with self-knowledge in the previous chat turns as a positive completion. For the negative prompt, we simply sample from the prompt with no additional prompting or few-shot turns.

Dataset	Type	Source	Number of Samples
UltraChat	SFT	HuggingFaceH4/ultrachat_200k	194,409
WizardLM	SFT	WizardLM/WizardLM_evol_instruct_V2_196k	80,662
SlimOrca	SFT	Open-Orca/SlimOrca-Dedup	143,789
ShareGPT	SFT	openchat/openchat_sharegpt4_dataset	3,509
Capybara	SFT	LDJnr/Capybara	7,291
Deita	SFT	hkust-nlp/deita-10k-v0	2,860
MetaMathQA	SFT	meta-math/MetaMathQA	394,418
Ultrafeedback	DPO	argilla/ultrafeedback-binarized-preferences	63,169
Orca Pairs	DPO	Intel/orca_dpo_pairs	12,859

Table 6: Fine-tuning datasets

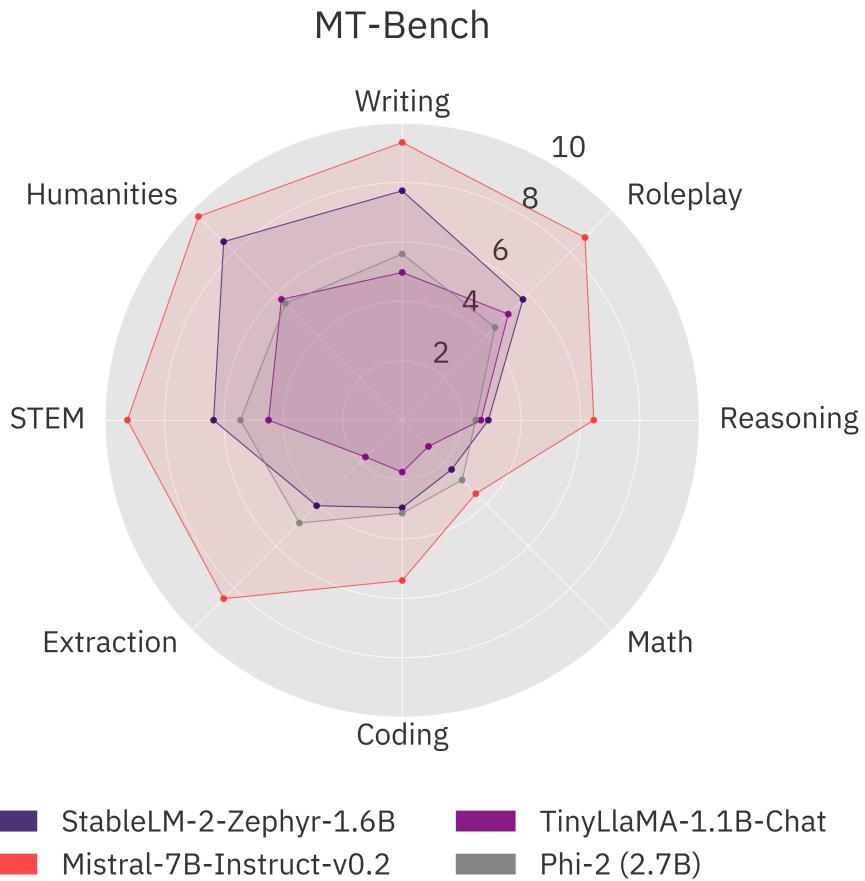


Figure 3: Stable LM 2 1.6B shows competitive performance, matching or even surpassing significantly larger models on MT-Bench.

We train with unpacked examples for 6 epochs using a batch size of 256, a warmup stage for 100 steps to a maximum LR of 3e-6 followed by a cosine decay to zero. The positive prompts are trained in the same way as the SFT step, while the negative prompts are trained with a negative token instead of the Assistant token.

Model	Size	MT-Bench
Mistral-7B-Instruct-v0.2	7B	7.61
Llama2-Chat	70B	6.86
stablelm-zephyr-3b	3B	6.64
MPT-30B-Chat	30B	6.39
stablelm-2-zephyr-1.6b	1.6B	5.42
Qwen-1.5-1.8B-Chat	1.8B	5.29
gemma-2b-it	2.5B	5.19
Falcon-40B-Instruct	40B	5.17
dolphin-2.6-phi-2	2.7B	4.93
phi-2	2.7B	4.29
TinyLlama-1.1B-Chat-v1.0	1.1B	3.46

Table 7: MT-Bench results

4 Experimental Results and Benchmarks

This section presents the main experimental results for Stable LM 2 1.6B. We compare with similarly-sized open-source models showing strong improvements on various tasks, including multilingual capabilities in Spanish, German, French, Italian, Portuguese, and Dutch. For context, we also provide comparisons with much larger models. We split our experiments into three sets: few- and zero-shot evaluations in English (as commonly done in the Hugging Face Open LLM leaderboard), multilingual evaluations, and conversational evaluations.

4.1 Few-shot and Zero-shot Evaluations

First, we assess the 0-shot and few-shot capabilities of Stable LM 2 by evaluating our model over popular benchmarks and comparing results against similarly sized open-source pre-trained models. Table 4 presents model evaluations in English. Our results cover the 6 benchmarks from the Open LLM Leaderboard ([5]): ARC-Challenge 25-shot [11] (**ARC**), HellaSwag 10-shot [83] (**HS**), MMLU 5-shot [30] (**MMLU**), TruthfulQA 0-shot [43] (**TQA**), WinoGrande 5-shot [62] (**Wino**) and GSM8K 5-shot [12] (**GSM**). Also, as Stable LM 2 is a general-purpose foundational model, we further assess natural language understanding capabilities by evaluating English and machine-translated versions of LAMBADA. All evaluations are performed with the Language Model Evaluation Harness framework [23][¶]. As shown in Table 4, Stable LM 2 1.6B (*stablelm-2-1-6b*) outperforms other base models by a significant margin. Similarly, the instruction-tuned version (*stablelm-2-1-6b-dpo*) improves on Microsoft’s Phi-1.5 by two average points while lagging behind the larger Phi-2.0 on few-shot accuracy. Performance versus Google’s Gemma 2B (2.5B parameters) is also remarkable.

4.2 Multilingual Evaluations

We assess knowledge and reasoning in the multilingual setting for non-English languages seen during pre-training by evaluating on ChatGPT-translated versions of ARC, HS, TQA, and MMLU ([38]). In addition, we test next-word prediction capabilities using the machine-translated LAMBADA datasets from [23]. After manual inspection by native speakers, we have deemed existing machine translations ** too noisy to draw accurate performance signals from. We instead evaluate multilingual next-word prediction with new translations which are made available for researchers ††.

The zero-shot results are presented in Tab. 5 and highlight Stable LM 2’s superior performance compared to models even twice its size.

[¶]<https://github.com/Stability-AI/lm-evaluation-harness/tree/stablelm-2/multilingual-bench>

^{**}https://huggingface.co/datasets/EleutherAI/lambada_openai

^{††}https://huggingface.co/datasets/marcob/lambada_multilingual

Framework	CPU	Precision	Throughput (Tok/s)	Power consumption (W)
MLX	M2	FP16	71	6
Mac Mini	(8GB)	INT4	127	11
GGUF	M2 Pro Max	FP16	46	14
2023 MacBook Pro	(16GB)	INT4	99	14

Table 8: Throughput and power usage on various devices using different quantization frameworks. We employ a batch size of 1 for all benchmarks. INT4 numbers for GGUF were collected using Q4_0 quantization.

4.3 MT Benchmark Evaluations

Finally, we also test the conversational skills of our model on the popular multi-turn benchmark **MT-Bench** [86]. The results are provided in Fig. 3 and Tab. 7. While lagging behind much more powerful models such as Mistral 7B Instruct v0.2 (more than 4x the size of Stable LM 2), our model delivers better chat performance and beats both Phi-2, Gemma 2B and TinyLLaMA 1.1B by a wide margin despite the larger size of the former.

5 Inference and Quantization

This model represents a substantial leap towards making advanced generation capabilities available directly on-device without the computational overhead of larger models. We believe this model strikes a great balance between remarkable efficiency and effectiveness in inference tasks when paired with inference frameworks and quantization methods. As part of our release, we provide quantized weights of *stablelm-2-1-6b* supported on popular inference libraries such as llama.cpp ^{††}, Apple MLX [29] and Intel OpenVINO ^{§§}.

5.1 Quantization

We make available quantization files for various models and formats to support easier integration with different inference frameworks, including:

- Two 4-bit quantized models: Q4_0, Q4_1 and a 5-bit quantized model: Q5_K_M GGUF
- INT4 for OpenVINO quantized with Intel’s Neural Network Compression Framework (NNCF)
- INT4 for MLX quantized with MLX

These quantization files can be found in the model’s Hugging Face repository for the convenience of developers and researchers working with our models. We aim to facilitate smoother deployment experiences across various deep-learning framework ecosystems by offering a range of quantization formats.

5.2 Throughput

In Tab. 8 we provide throughput numbers obtained from our model running on consumer-grade devices and the system environments utilized. Our initial runs showcase that when using a lower precision, we are able to achieve almost 2x performance in throughput. Note that these figures are provided as a reference, and they are not the result of rigorous benchmarking but are rather intended to give users a practical insight into what they can expect in terms of performance on commonly used devices. Likewise, as lower precision quantization is expected to reduce the model’s performance, we encourage researchers and developers to assess the potential degradation in real-world scenarios.

^{††}<https://github.com/ggerganov/llama.cpp>

^{§§}https://github.com/openvino_toolkit/openvino

6 Future Work

There is a number of research avenues we would like to explore to further improve the model:

1. **Data.** In this work, we focused on publicly available data. In particular, most of the data comes from web-crawled content, as is common for most models. This data is known to contain many low-quality documents [20] that can potentially harm training. We believe there is significant potential in smart filtering, re-writing, and synthetic data generation with strong models.
2. **Hallucination Mitigation.** Language models are prone to generating incorrect or misleading information, and small language models are even more prone to doing so. Finding reliable ways to detect hallucinations in these models will unlock new use cases in areas that are sensitive to hallucinations.
3. **Long Contexts and Retrieval.** The ability to retrieve information across long context windows is essential for applications such as chat models or dataset integration. Accordingly, in Appendix C, we explore the current capabilities and limitations of StableLM2 1.6B on the Needle-in-the-Haystack task. Going forward we plan to further build upon this work as well as to extend our models to context lengths beyond 4k.
4. **Conditional Computation.** Small models are often capacity-constrained – that is, with the current training approaches, they lack the capacity to process and exploit all of the training data. Recently, ideas such as Mixture of Experts have been successfully applied to take a dense model and extend it to contain more parameters that are selectively applied to certain inputs (for instance, via sparse upcycling [37]). Importantly, if each token selects only one expert, the overall inference FLOPs do not significantly change. Applying this to the Stable LM 2 1.6B model is a natural extension we will investigate.

7 Environmental and Societal Impact

7.1 Carbon Footprint

The training of Stable LM 2 has consumed energy with associated carbon dioxide emissions. In line with [67] we report our carbon footprint based on the formula

$$\text{Total Wh} = \text{GPU-h} \times \text{power consumption} \times \text{PUE}$$

where the Power Usage Effectiveness is set to 1.1. We trained Stable LM 2 for $\approx 92,000$ GPU-hours, giving a total power consumption of 30MWh considering our average power usage. The tons of emitted carbon tCO₂eq can be estimated using the US national average carbon intensity factor of 0.385 kg CO₂eq/KWh, leading to a final figure of 11 tCO₂eq.

7.2 Societal impact

Stability AI is committed to releasing open models to help improve access to foundational AI technology. Open access to model weights enables researchers to inspect models for suitability and vulnerabilities, test the effectiveness of different optimization strategies, and correct for biases observed in the model. To that end, this model is released under an open non-commercial license. However, open release can introduce challenges in assessing the societal impact of a model. For example, Stability AI does not have direct visibility into downstream applications of Stable LM 2 1.6B, the distribution of applications by sector, or the distribution of model usage by geography. Since the model is released with a noncommercial license, we expect a limited number of applications outside of fine-tuning or evaluation interfaces and a limited number of third parties affected by the model.

We will continue to monitor openly released fine-tuned models to understand the extent of fine-tuning research or development activity that uses Stable LM 2 1.6B as a base model, including the evaluation results of these derivative models.

8 Conclusion

In this report, we introduced Stable LM 2 1.6B, a compact decoder-only language model trained on multilingual datasets. It fluidly handles up to seven languages: English, Spanish, German, Italian, French, Portuguese, and Dutch. To ensure that the community can reproduce our run, we detail all datasets used during training –with the *exact* data mix– and our newly designed learning rate schedule. We also conduct extensive model evaluations and comparisons with other similarly-sized models, demonstrating Stable LM 2 1.6B’s exceptional performance. Finally, we profile the model on common edge computing architectures. We hope the current report contributes to the improvement and further research on small language models.

Acknowledgments

We thank our awesome MLOps team members, particularly Richard Vencu, for the support provided. We also thank Christian Laforte, Cedric Wagrez, and Jerry Chi for their feedback, useful ideas, and comments.

References

- [1] Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- [2] b-mc2. sql-create-context dataset, 2023.
- [3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.
- [4] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.
- [5] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard, 2023.
- [6] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. Gpt-neox-20b: An open-source autoregressive language model, 2022.
- [7] British Library, Victoria Morris, Daniel van Strien, Giorgia Tolfo, Lora Afric, Stewart Robertson, Patricia Tiney, Annelies Dogterom, and Ildi Wollner. 19th century books - metadata with additional crowdsourced annotations, 2021.
- [8] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S. Weld. TLDR: Extreme summarization of scientific documents. *arXiv:2004.15011*, 2020.
- [9] Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, mar 2020. Data available at <https://github.com/PolyAI-LDN/task-specific-datasets>.
- [10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.
- [11] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.
- [12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.
- [13] Together Computer. Redpajama: An open source recipe to reproduce llama training dataset, 2023.
- [14] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.
- [15] Luigi Daniele and Suphavadeeprasit. Amplify-instruct: Synthetically generated diverse multi-turn conversations for efficient llm training. *arXiv preprint arXiv:(coming soon)*, 2023.
- [16] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning, 2023.
- [17] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.
- [18] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.

- [19] Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellan, Kevin Robinson, Kathleen Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts, 2022.
- [20] Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hanna Hajishirzi, Noah A. Smith, and Jesse Dodge. What’s in my big data?, 2023.
- [21] Yao Fu, Rameswar Panda, Xinyao Niu, Xiang Yue, Hannaneh Hajishirzi, Yoon Kim, and Hao Peng. Data engineering for scaling language models to 128k context, 2024.
- [22] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020.
- [23] Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, September 2021.
- [24] Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. Openwebtext corpus, 2019.
- [25] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour, 2018.
- [26] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [27] Himanshu Gupta, Neeraj Varshney, Swaroop Mishra, Kuntal Kumar Pal, Saurabh Arjun Sawant, Kevin Scaria, Siddharth Goyal, and Chitta Baral. “john is 50 years old, can his son be 65?” evaluating NLP models’ understanding of feasibility. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 407–417, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.
- [28] Johannes Hagemann, Samuel Weinbach, Konstantin Dobler, Maximilian Schall, and Gerard de Melo. Efficient parallelization layouts for large-scale distributed model training, 2023.
- [29] Awni Hannun, Jagrit Digani, Angelos Katharopoulos, and Ronan Collobert. MLX: Efficient and flexible machine learning on apple silicon, 2023.
- [30] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.
- [31] 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. Training compute-optimal large language models, 2022.
- [32] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. Camels in a changing climate: Enhancing lm adaptation with tulu 2, 2023.
- [33] Mingi Jeon, Seung-Yeop Baik, Joonghyuk Hahn, Yo-Sub Han, and Sang-Ki Ko. Deep Learning-based Code Complexity Prediction, 2022.
- [34] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *arXiv e-prints*, page arXiv:1705.03551, 2017.
- [35] Greg Kamradt and Ikko Ashimine. Needle in a haystack - pressure testing llms, 2023.
- [36] Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.
- [37] Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *arXiv preprint arXiv:2212.05055*, 2022.
- [38] Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A Rossi, and Thien Huu Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. *arXiv preprint arXiv:2307.16039*, 2023.

- [39] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zhetlonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. Starcoder: may the source be with you!, 2023.
- [40] Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zhetlonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. StarCoder: may the source be with you!, 2023.
- [41] Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification, 2023.
- [42] Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, November 2020. Association for Computational Linguistics.
- [43] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods, 2022.
- [44] Emmy Liu, Chen Cui, Kenneth Zheng, and Graham Neubig. Testing the ability of language models to interpret figurative language, 2022.
- [45] Tiedong Liu and Bryan Kian Hsiang Low. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks, 2023.
- [46] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning, 2023.
- [47] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*, 2023.
- [48] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023.
- [49] Risto Luukkonen, Ville Komulainen, Jouni Luoma, Anni Eskelinen, Jenna Kanerva, Hanna-Mari Kupari, Filip Ginter, Veronika Laippala, Niklas Muennighoff, Aleksandra Piktus, Thomas Wang, Nouamane Tazi, Teven Le Scao, Thomas Wolf, Osma Suominen, Samuli Sairanen, Mikko Merioksa, Jyrki Heinonen, Aija Vahtola, Samuel Antao, and Sampo Pyysalo. Fingpt: Large generative models for a small language, 2023.
- [50] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.
- [51] Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018.
- [52] John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. English WordNet 2019 – an open-source WordNet for English. In Piek Vossen and Christiane Fellbaum, editors, *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland, July 2019. Global Wordnet Association.
- [53] Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. Scaling data-constrained language models, 2023.

- [54] Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages, 2023.
- [55] OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yafei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiro, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotstet, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.
- [56] Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text, 2023.
- [57] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023.
- [58] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.
- [59] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.
- [60] Vipul Raheja, Dhruv Kumar, Ryan Koo, and Dongyeop Kang. Coedit: Text editing by task-specific instruction tuning, 2023.
- [61] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020.
- [62] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

- [63] Eva Sharma, Chen Li, and Lu Wang. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. *CoRR*, abs/1906.03741, 2019.
- [64] Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11321–11329, Jun. 2022.
- [65] Gizem Soğancioğlu, Hakime "Ozt"urk, and Arzucan "Ozg"ur. Biosses: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58, 2017.
- [66] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.
- [67] 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. Llama: Open and efficient foundation language models, 2023.
- [68] Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. Stablelm 3b 4e1t, 2023.
- [69] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Shengyi Huang, Kashif Rasul, Alexander M. Rush, and Thomas Wolf. The alignment handbook. <https://github.com/huggingface/alignment-handbook>, 2023.
- [70] Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.
- [71] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [72] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data, 2023.
- [73] Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Polak Scowcroft, Neel Kant, Aidan Swope, and Oleksii Kuchaiev. Helpsteer: Multi-attribute helpfulness dataset for steerlm, 2023.
- [74] Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities, 2023.
- [75] Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard H. Hovy. Large-scale cloze test dataset designed by teachers. *CoRR*, abs/1711.03225, 2017.
- [76] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023.
- [77] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, 2023.
- [78] Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. Rlcld: Reinforcement learning from contrast distillation for language model alignment, 2023.
- [79] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2020.
- [80] Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.
- [81] Weizhe Yuan and Pengfei Liu. restructured pre-training, 2022.
- [82] Armel Randy Zebaze. Self-instruct-starcoder.
- [83] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics.
- [84] Biao Zhang and Rico Sennrich. Root mean square layer normalization, 2019.
- [85] Jingmiao Zhao and Carolyn Jane Anderson. Solving and generating npr sunday puzzles with large language models, 2023.

- [86] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.
- [87] Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. Temporal reasoning on implicit events from distant supervision. In *NAACL*, 2021.
- [88] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. Lima: Less is more for alignment, 2023.

Source	Sampling weights					
	Control	Mix 1	Mix 2	Mix 3	Mix 4	Mix 5
Cultura-x En	0.6894	0.5694	0.3294	0.1494	0.49	0.49
Cultura-x De	0.	0.03	0.06	0.09	0.	0.
Cultura-x Es	0.	0.03	0.06	0.09	0.	0.
Cultura-x Fr	0.	0.03	0.06	0.09	0.	0.
Cultura-x It	0.	0.03	0.06	0.09	0.	0.
Cultura-x Pt	0.	0.03	0.06	0.09	0.	0.
Cultura-x Nl	0.	0.03	0.06	0.09	0.	0.
Starcoder	0.0702	0.0702	0.0702	0.0702	0.0497	0.28
Others	0.2292	0.2292	0.2292	0.2292	0.4648	0.2292

Table 9: Data ablations with corresponding sampling weights. Column 2 (**Control**) is our reference, containing only English and code data, with a standard, significant amount of web data. **Mix 1-3** test our strategy for adding multilingual data, capping non-English sources to a fixed percentage. In **Mix 4-5** we reduce the amount of web data, increasing respectively books and academic sources and code. Source **Others** contains the same data as in Tab. 1 from the categories: Academic, Books, Social, Law, Math, and Wiki. In **Mix 4**, we uniformly upsample Academic and Books sources.

A Data Ablations

How to select the optimal training mix for pre-training from a set of sources is an open problem. Tuning weights based on downstream tasks [10, 19] can be extremely costly and bears the risk of overfitting on particular tasks as well as exploiting data leakage. While the computationally cheap, principled approach introduced in [76] is promising, we found it delivers sub-optimal weights when the data sources are highly imbalanced and have different information content (e.g., large web sources vs curated datasets). Furthermore, multilingual evaluations introduce a more explicit dependence on the tokenizer, with increased noise due to the lack of high-quality, non-machine-translated benchmarks. We, therefore, aim to find general guiding principles that are expected to hold against changes in the tokenizer or in the absence of high-quality benchmarks for each data category while keeping the cost of these ablations low.

We trained a set of 1B models on a total of 100B tokens sampled according to Tab. 9. Evaluations of each model on English and non-English benchmarks are shown in Tab. 10. We observe the following trends

- Contrary to [53], we find less conclusive evidence that code can be used as a neutral filler for training data as increasing the amount of code leads to a degradation in language model performances. We leave a more thorough exploration of this to future work, including math and reasoning tasks that may benefit from a higher proportion of code data.
- Performance on non-English benchmarks increases for each language by adding any amount of data in that same language. However, this increase saturates very fast and we observe only modest gains beyond 6%. We hypothesise that this might be due to the lack of high-quality, structured data sources in non-English languages, which we only sample from the web.
- Upsampling Academic and Books sources improves downstream performance over the control run, particularly in natural language understanding.

B Scheduler Ablations

[31] shows that the widely adopted cosine learning rate decay achieves optimal performance only when performing a full cosine period, forcing practitioners to fix the number of steps beforehand. As multi-epoch training performs well for LLMs [53, 49, 68], and larger and cleaner data sources are

Data	Avg	English Commonsense	LAMBADA	Okapi
Control	38.76	66.51	30.93	33.98
Mix 1	39.91	63.92	34.25	35.39
Mix 2	40.69	64.71	35.74	35.93
Mix 3	39.87	63.02	35.22	35.25
Mix 4	39.41	65.87	32.73	34.58
Mix 5	38.66	64.95	31.4	34.07

Table 10: Downstream evaluations of the models considered in our data ablations. English commonsense includes: **ARC-Easy**, **PIQA**, **SciQ**, **WinoGrande**. Each score is averaged over EN, DE, ES, FR, and IT.

made accessible by the OS community, it becomes more and more important to alleviate this limitation. To this end, we experimented with the "inverse square root" (rsqrt) learning rate scheduler [59] Eq. 2

$$\frac{1}{\sqrt{\max(i, k)}} \quad (2)$$

where i is the current iteration and k the number of warmup steps. As the scheduler is strictly convex and reaches zero asymptotically, it can be used to train for infinite iterations.

However, in standard scenarios, of which we show an example in Fig. 4, rsqrt consistently underperforms cosine. We make the comparison by decaying the learning rate to 0 in both cases, with a linear cool down for the last 10% of the steps for the rsqrt scheduler.

A sharp difference between the two schedulers is how they start from the peak learning rate: with a flat slope, negative second derivative for cosine and a large slope, positive second derivative for rsqrt. This allows rsqrt to escape the high learning rate region quickly, which we identified as the main cause of the performance gap. We then get the best of both worlds by combining both schedulers into a single function that is again the cosine for the first section, and smoothly switches to rsqrt, as described in Eq. 1.

Finally, we experimented with two different versions of our scheduler, whose performance we show in the right panels of Fig. 4: hybrid (**hyb**) corresponds to our final scheduler defined in Eq. 1, while for hybrid₂ (**hyb**₂) we double the cosine period and move the turning point at half the training steps instead of at a fourth. We attribute the difference in performance between the two versions to the different average learning rates. For the set of experiments in Fig. 4, integrating the scheduler over the training volume, we obtain an average learning rate of $1.5e-4$, $1.13e-4$ and $1.67e-4$ for cosine, hyb and hyb₂ respectively. We leave to future work a proof that schedulers with the same average value lead to statistically equivalent models under mild conditions, such as monotonicity and fixed endpoints.

C Evaluation of Performance Across Different Sized Context Windows

The Needle-in-a-haystack test, as introduced in [35], is commonly used to assess the retrieval capabilities of LLMs across different context window sizes. Following the methodology of [35], a fact ("the needle") is embedded within a context of unrelated essays ("the haystack"). The model under evaluation is tasked to answer a question requiring the retrieval of the needle from the context. The evaluation is carried out systematically by placing the needle at 35 different depths within the context and comparing the results across context window sizes from 500 to 4000 tokens. Notably, these window sizes are specifically chosen to assess the performance within our trained context size of 4096 tokens. An AI judge, typically GPT-4 [55], scores the answers from 1 to 10 based on whether or not the fact was correctly retrieved.

While performing the evaluation, we noticed that the order of the context was not fixed or controlled by a seed, leading in some cases to significantly different scores. For this reason, we show in Fig. 5

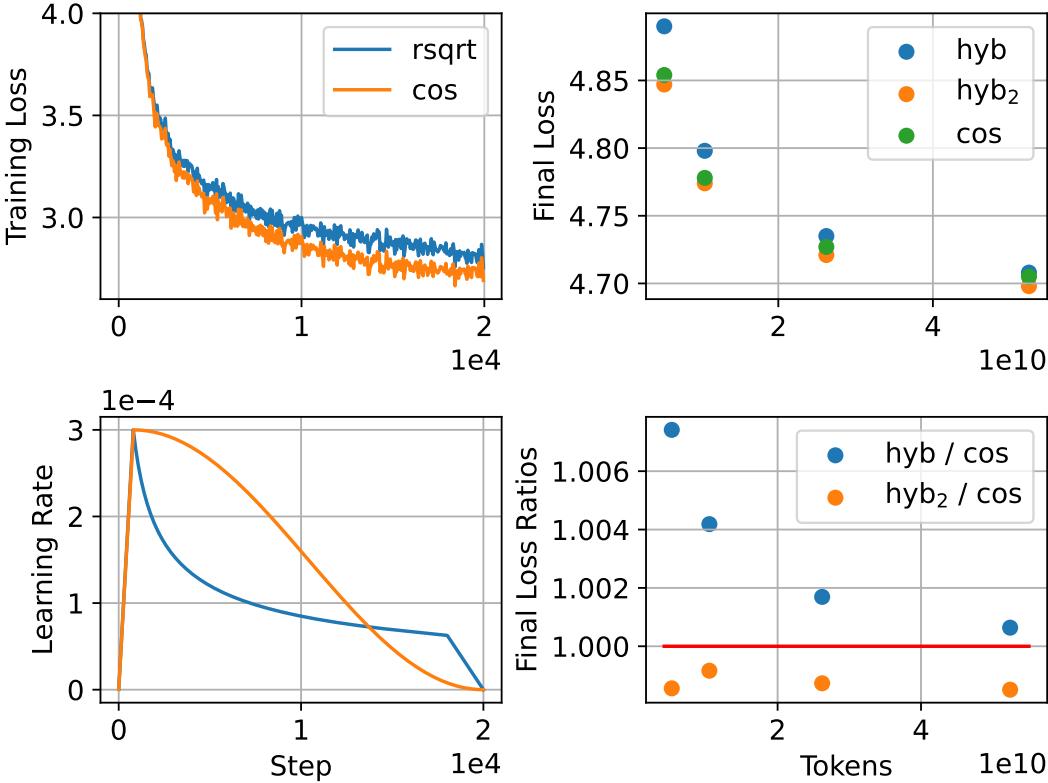


Figure 4: Learning rate scheduler ablations. Left figures: comparison between rsqrt and cosine decay scheduler. Right figures: final loss of models trained with the two variants of the hybrid scheduler and cosine for different number of training tokens (top). The difference in final loss between the 3 schedulers is < 1% (bottom).

the average results over 10 different runs, together with the corresponding standard deviation. For Stable LM 2, we employ the prompt of [21], whereas for our fine-tuned version we use the official repository as is. Averaging the scores over the evaluated grid, we observe a slight degradation from ≈ 7.7 for our base model to ≈ 6.8 for the fine-tuned version. The different prompt structures make it hard to directly compare the results, however, we will investigate in future work how different elements such as the distribution of the document lengths and the attention mask correlate with this behavior.

D Global Batch Size

The role of the batch size in stochastic optimization convergence is illustrated in [51] as follows. The loss $L(\theta)$ of a model parameterized by θ over a training set D is estimated by independently drawing random samples from D to form a batch B . Thus, the loss gradient used to update the model's parameters is given by

$$\nabla L(\theta) = \frac{1}{|B|} \sum_{i=1}^{|B|} \nabla_{\theta} L_{x_i}(\theta) \quad (3)$$

and its variance scales with $1/|B|$. In other words, a bigger batch has a lower variance and thus provides a more accurate estimate of the gradient. A more accurate gradient suggests that we should increase the learning rate accordingly to converge faster; however, in practice, this is far from trivial and requires special handling such as fine-tuning the learning rate scheduler or performing per-layer updates [25, 79].

To choose the global batch size for Stable LM 2, we make the following assumptions:

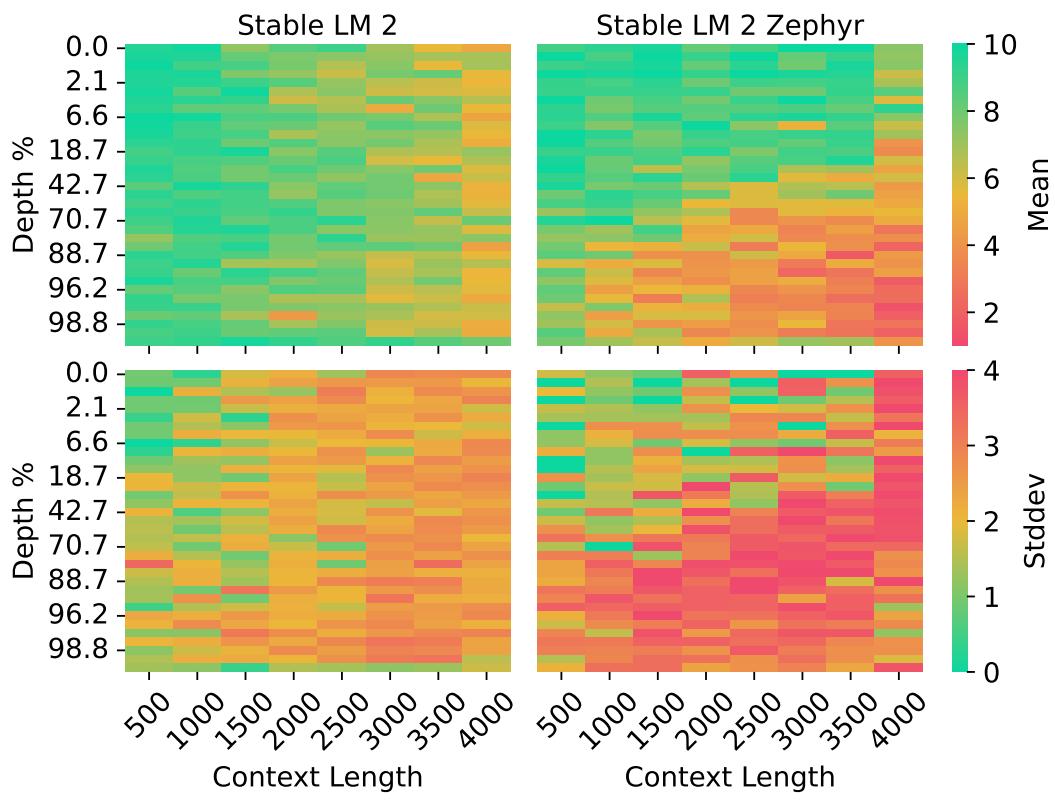


Figure 5: Needle-in-a-haystack evaluation of Stable LM 2 on context window sizes from 500 to 4000 tokens.

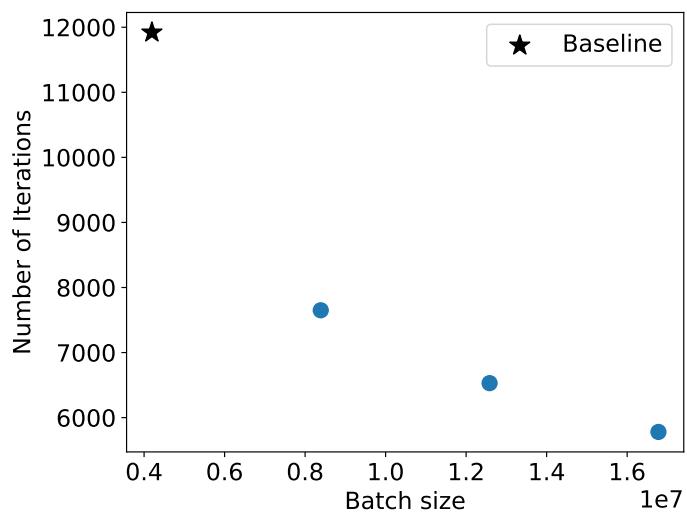


Figure 6: Number of iterations performed at various batch sizes to match the same final loss.

1. We do not change the learning rate based on the batch size
2. The batch size can be approximately scaled at no cost.

With 1., we are, in principle, giving up on potential further gains by tuning the step size to the gradient noise. In practice, we notice that using a large learning rate at the boundary between convergence and divergence of our ablations is more than enough to compensate for this. 2. follows from a combination of theoretical results, hardware optimizations, and increased training data availability. [53] empirically demonstrated how multiple epochs of repeated data are as good as fresh data in LLMs pre-training, while the computational overhead of increasing data parallel workers is minimal.

The data volume available for pre-training is consistently increasing through larger datasets, and there are promising results of multi-epoch training. Hence, we explore training with larger batch sizes, which requires more overall training tokens to reach the final loss but significantly decreases training time. To determine our final batch size, we start by training a baseline with a batch B_0 of 4M tokens on $T_0 = 50B$ total training tokens. Subsequently, we train new models from scratch with batches $B_i = 8M, 12M, \text{ and } 16M$, increasing T_i tokens until the same final loss is reached. We do this with an rsqrt scheduler as the number of required training steps is unknown beforehand.

In Fig. 6, we show the number of iterations B_i/T_i which are required to match the baseline loss. Assuming for simplicity that the iteration time is independent of the batch size, this gives an upper bound on the speed up we can achieve by increasing the batch.

In the regime considered, we observe that an increase in batch size leads to a decrease in the number of iterations required to match the baseline loss all the way to the biggest batch considered of 16.7M tokens, which speeds up training by a factor 2x. However, to achieve an equal loss, we require $T_i = 96B$ training tokens, which is a factor 1.96 increase compared to the baseline. Therefore to train Stable LM 2, we opt for a global batch of 8,388,608, which with the layout of Tab. 3 offers the best compromise between decrease in training time and additional required training tokens.

E Restructured Pre-training Sources

While many sources are restructured in [81], in this work, we have only considered those listed in Tab.11 with commercially viable licenses such as BIGPATENT, CLOTH, SciTLDR, TriviaQA, WordNet, WikiHow, etc. We also restructure additional sources with similar licensing compatibility and list them below.

Dataset	Prefix
Banking77 [9]	banking77
BigPatent [63]	big_patent
BIOSSES [65]	biosses
BLBooksGenre [7]	TheBritishLibrary/blbooksgenre
CodeComplex [33]	codeparrot/codecomplex
CoEdIT [60]	grammarly/coedit
CLOTH [75]	AndyChiang/cloth
CommonGen [42]	common_gen
FigQA [44]	nightingal3/fig-qa
FeasibilityQA [27]	jon-tow/feasibility_qa
Flan 2021 [47]	DataProvenanceInitiative/flan2021_submix_original
Flan Chain of Thought [47]	DataProvenanceInitiative/cot_submix_original
Flan NIV2 [47]	DataProvenanceInitiative/niv2_submix_original
Flan T0 [47]	DataProvenanceInitiative/t0_submix_original
HelpSteer [73]	nvidia/HelpSteer
IMDB Review [50]	ajaykarthick/imdb-movie-reviews
Joke Explanation [47]	dim/joke_explaination
MBPP [1]	mbpp
NarrativeQA [36]	narrativeqa
PuzzleQA [85]	Jingmiao/PUZZLEQA
SciTLDR [8]	allenai/scitldr
Self-Instruct Starcoder [82]	codeparrot/self-instruct-starcoder
SQL Create Context [2]	b-mc2/sql-create-context
StepGame [64]	tasksource/stepgame
TRACIE [87]	tasksource/tracie
TriviaQA [34]	trivia_qa
WikiHow	wikihow
WordNet [52]	jon-tow/open-english-wordnet-synset-2023
Yahoo Answers Topics	yahoo_answers_topics

Table 11: The original sources for restructured and instruction pre-training datasets can be found at <https://huggingface.co/datasets/> followed by the provided prefix.