Rephrasing natural text data with different languages and quality levels for Large Language Model pre-training

Michael Pieler **Marco Bellagente** **Hannah Teufel** **Duy Phung**
Nathan Cooper Jonathan Tow Paulo Rocha Reshinth Adithyan
Zaid Alyafeai Nikhil Pinnaparaju Maksym Zhuravinskyi Carlos Riquelme

Stability AI Language Team*

Abstract

Recently published work on rephrasing natural text data for pre-training LLMs has shown promising results when combining the original dataset with the synthetically rephrased data. We build upon previous work by replicating existing results on C4 and extending them with our optimized rephrasing pipeline to the English, German, Italian, and Spanish Oscar subsets of CulturaX. Our pipeline leads to increased performance on standard evaluation benchmarks in both the mono- and multilingual setup. In addition, we provide a detailed study of our pipeline, investigating the choice of the base dataset and LLM for the rephrasing, as well as the relationship between the model size and the performance after pre-training. By exploring data with different perceived quality levels, we show that gains decrease with higher quality. Furthermore, we find the difference in performance between model families to be bigger than between different model sizes. This highlights the necessity for detailed tests before choosing an LLM to rephrase large amounts of data. Moreover, we investigate the effect of pre-training with synthetic data on supervised fine-tuning. Here, we find increasing but inconclusive results that highly depend on the used benchmark. These results (again) highlight the need for better benchmarking setups. In summary, we show that rephrasing multilingual and low-quality data is a very promising direction to extend LLM pre-training data.

1 Introduction

The rapid advancements of Large Language Models (LLMs) have transformed them into powerful tools for natural language processing (NLP). While early breakthroughs were often driven by architectural innovations, the field has recently converged toward similar Transformer-based architectures. As a result, differences in the training data composition have emerged as a critical differentiator to create high-performance LLMs. Furthermore, as model pre-training continues to be scaled, it has been questioned if natural text data may become scarce [1].

Synthetic data has been proposed to alleviate potential data bottlenecks, while offering increased control over the data properties. By leveraging a large and diverse set of web-based text, such approaches could mitigate biases from the model or task they are based on. (See related work in section 2.)

*Correspondance to: michael@stability.ai

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

In this work, we build upon the text rephrasing from Maini, et al. (2024) [2]. Their rephrasing setup leverages LLMs to create large scale pre-training data. Such a rephrasing setup can be interpreted as a form of paraphrasing text augmentation [3], that is applied to longer text passages. Those passages are rewritten in different styles by prompting an LLM with specific instructions. As a first step, we focus on replicating the work from Maini, et al. (2024) [2] on rephrasing the English Colossal Clean Crawled Corpus (C4) dataset [4]. We then extend this approach to additional datasets with different languages and varying perceived quality levels. Our contributions can be summarized as follows:

- We successfully replicate the rephrasing study by Maini, et al. (2024) [2] with a larger LLM trained with double the batch size on C4.
- We successfully extend our rephrasing pipeline to the English, German, Italian, and Spanish Oscar subsets of CulturaX (CX) [5].
- In addition, we compare the rephrasing setup on web-based datasets with different perceived quality levels, i.e., CX, C4, and FineWeb-Edu (FWE) [6].
- Furthermore, we study the impact of the rephrasing model, by comparing two state-of-the-art LLMs, namely Mistral 7B Instruct v0.2 [7] and Qwen2 7B Instruct [8] as well as differently sized models from the Qwen2 family.
- Finally, we apply supervised fine-tuning (SFT) to our LLMs pre-trained on rephrased data to investigate its downstream effect.

2 Related work

Curating and selecting data for LLM pre-training is an active area of research, and spans a wide range of approaches and techniques [9, 10]. Different pipelines have shown how to prepare web data into training data [11, 12, 13, 14], including heuristics based filtering [15], classifiers trained to select content with high educational value [16, 13], and de-duplication [17, 18]. At the granularity of a fixed, given corpus Sang, et al. (2023) [19] proposed to compute weights based on a distributionally robust proxy model. Furthermore, a recent setup uses pre-trained open source models to estimate correlations between benchmark scores and documents [20].

A large variety of text data augmentation strategies have been described for NLP [3]. Of particular interest for our work is paraphrasing. Paraphrasing can be carried out at different levels, i.e., starting at a level of words, phrases, sentences, up to text passages, and with different setups, ranging from simpler setups, e.g., Thesauruses- [21] or rule-based [22, 23], up to more complex setups, e.g., by using embeddings [24] or generating text with LLMs [25].

Recently, generating synthetic training data with LLMs has been increasingly explored. For pre-training data a range of pipelines have been published to generate, e.g., synthetic short stories [26], textbooks and exercises [27], blog posts, stories, posts, and other articles [28], as well as web-based text rephrased in different styles [2]. Additionally, several setups have been described for fine-tuning, e.g., back translation of instruction data [29], creation of alignment data [30, 31, 32], generating text data based on specific taxonomies [33], use cases, skills [34, 35], or personas [36]. Furthermore, synthetic text data can also be useful for inference, e.g., for test-time augmentation [37].

In addition, we point the interested reader to more extensive reviews of this topic [38, 39, 40, 41].

3 Methods

3.1 Natural text data

For our experiments, we compare three web-based pre-training datasets: C4 [4], the Oscar subsets from CulturaX (CX) [5], and the FineWeb-Edu (FWE) dataset [6]. These datasets undergo different processing and filtering heuristics, enabling us to validate our method on

datasets of varying perceived quality and quantify the potential improvements achieved by rephrasing thoroughly cleaned data All natural text data sources are listed in Table 1 with their document and token counts for the original and the rephrased datasets. The token counts are based on the Stable LM 2 tokenizer [42]. The Ask-LLM [43] subsets are filtered with a score threshold of >0.6 (A>0.6) and >0.97 (A>0.97), and the FWE classifier subsets are filtered with a score threshold of >2.5 (F>2.5; for more details about the classifiers see Appendix D).

Table 1: Overview of all original and rephrased (re.) datasets. C4, CulturaX (CX), and FineWeb-Edu (FWE) are used as base datasets. The default rephrasing model is Mistral 7B Instruct v0.2, except for the Qwen2 Instruct model size experiments. The classifier filtered datasets are indicated with the corresponding threshold, e.g., Ask-LLM > 0.6 (A>0.6), FWE > 2.5 (F>2.5). The token counts are based on the Stable LM 2 tokenizer.

Dataset	**mio. docs**	**B tokens**
C4 (English)	365	172
↳ **C4 A>0.6**	34	20
C4 rephrased toddler	343	85
↳ **C4 rephrased toddler A>0.6**	336	84
↳ **C4 rephrased toddler A>0.97**	33	5
C4 rephrased hard	341	167
↳ **C4 rephrased hard A>0.6**	338	160
↳ **C4 rephrased hard A>0.97**	38	18
C4 rephrased wiki	334	161
↳ **C4 rephrased wiki A>0.6**	332	152
↳ **C4 rephrased wiki A>0.97**	52	16
C4 rephrased QA	299	176
↳ **C4 rephrased QA A>0.6**	298	162
↳ **C4 rephrased QA A>0.97**	23	9
CX-E (English)	843	741
CX-E rephrased toddler	227	93
CX-E rephrased wiki	170	130
CX-E rephrased QA	157	137
CX-E Qwen2 0.5B Inst. re. QA opt.	104	73
CX-E Qwen2 1.5B Inst. re. QA opt.	107	84
CX-E Qwen2 7B Inst. re. QA opt.	105	102
CX-G (German)	105	96
CX-G rephrased QA optimized	85	56
CX-S (Spanish)	86	78
CX-S rephrased QA optimized	75	51
CX-I (Italian)	46	48
CX-I rephrased QA optimized	38	28
FWE	97	100
FWE rephrased QA optimized	95	98
↳ **FWE re. QA opt. F>2.5**	74	79
Total original	1,542	1,234
Total unique rephrased	2,777	1,630

3.2 Rephrasing natural text data

The rephrasing pipeline consists of several steps: In the preprocessing step, we first split the raw text into smaller passages (section 3.2.1). These passages are then inserted into prompt templates (section 3.2.2) for model inference to obtain the rephrased text (section 3.2.3). The rephrased passages are cleaned and merged back into full documents again in the postprocessing step (section 3.2.4). Finally, the rephrased data is used for the LLM pre-training (Appendix E; pipeline overview in Appendix A).

3.2.1 Preprocessing

The preprocessing splits the documents into smaller passages. We extend the passage length of 300 tokens from Maini, et al. (2024) [2] to 350 tokens and introduce a minimum passage length of 50 tokens. We increased the token limit to 350 due to our merge logic and did not observe any degradation in our rephrasing experiments. For passage generation, we use a simple and fast preprocessing algorithm:

1. Generate smaller passages, by splitting documents on line breaks, i.e., `\n`.
2. Remove empty passages.
3. Split passages exceeding 350 tokens on common sentence-ending characters, i.e., `.`, `!`, or `?` followed by one or more white space characters.
4. Merge consecutive passages until we reach the maximum target length of 350 tokens. If the total passage length is longer than 350 tokens after a merge we start a new merge process with the last passage.

To obtain the token counts in the preprocessing step, we use a token count per character estimate based on a random subset of the data to avoid running tokenization every time we want to verify the passage length. Compared to the pre-processing setup described by Maini, et al. (2024) [2], we use simple regex patterns instead of the NLTK sentence splitter and omit the check for specific delimiters in the text. However, our results on C4 are better than theirs and are discussed in detail in section 4.1.

3.2.2 Prompt templates

In this work we use two different sets of prompt templates. First, for replicating the previous work on C4 (section 4.1) and comparing to the CX English (CX-E) dataset (section 4.2), we use the toddler, hard, wiki, and QA prompt templates from Maini, et al. (2024) [2]. We adapted the toddler, hard, wiki, and QA prompt templates after initial experiments with the newer model version v0.2 of the Mistral 7B Instruct model. Then, we create new prompt templates, optimized for easier postprocessing, which are applied in the experiments on the FWE dataset in the data quality investigations (section 4.3) and the model scale ablations (section 4.4). Finally, we translate those prompt templates to other languages, i.e., German, Spanish, and Italian, for our multilingual experiments (section 4.2). All prompt templates are shown in Appendix B.

3.2.3 Inference

We rely on Mistral 7B Instruct v0.2 [7] for inference, in contrast to the setup used by Maini, et al.(2024) [2], which used Mistral 7B Instruct v0.1. In order to investigate the different performance of similarly sized models, as well as differently sized models within the same family, we additionally use the Qwen2 0.5B, 1.5B, and 7B Instruct models [8] for QA rephrasing the CX-E dataset in section 4.4.

3.2.4 Postprocessing

In the postprocessing, we apply two different transformations based on whether the outputs were generated with prompts following Maini, et al. (2024) [2] or our optimized prompt templates for simple text extraction. Below are the postprocessing steps for the outputs generated using prompts based on Maini, et al. (2024) [2]:

1. Identify and split multiple paraphrases within the output and randomly return one.
2. Remove any unwanted elements in the output, such as `Paraphrase:`, `Toddler-friendly paraphrase:`, `Erudite paraphrase:`, and other patterns.
3. Only keep passages between 50 and 5,000 characters.
4. Discard passages if the last character is alphabetic, as this indicates a truncated output.
5. Merge passages to the full documents again.
6. Remove documents with fewer than 100 characters.

For outputs generated with our optimized prompt setup, we replace step 1 and 2 from above by extracting the content between the first `<text></text>` tag pair, which is a much simpler and more efficient setup.

4 Results

We evaluate our pre-training experiments with the Language Model Evaluation Harness [44] (MIT License) on the following established natural language benchmarks: ARC Challenge (A-C), ARC Easy (A-E) [45], HellaSwag (HS) [46], Lambada (L) [47], PIQA (P) [48], SciQ (SQ) [49], and WinoGrande (WG) [50]. For the evaluation of our fine-tuned models, we use the Open LLM 1 and 2 benchmarks [51, 52]

4.1 Rephrasing C4

First, we rephrase C4 to verify that we see similar results with our experimental setup compared to previous work by Maini, et al. (2024) [2]. Table 2 shows our C4 rephrasing experiments grouped together with the most similar experimental setup from Maini, et al. (2024) [2]. Our setup outperforms the previous work between 0.2 and 2.5 percentage points. Interestingly, we observe larger, systematic differences at the benchmark level. In our experiments, scores on A-C and HS are always higher, whereas on A-E they are always lower. Those differences can be very likely linked to the different training setups and models used for rephrasing the data (see Appendix E and section 3.2.3).

Table 2: Replicated C4 experiments with the baseline and rephrased (re.) data. Experiments are grouped together with the most similar experiment from Maini, et al. (2024) [2]. Each section uses the experiment from Maini, et al. (2024) [2] as baseline to indicate the change. The used rephrasing prompt is indicated in the dataset name with hard, QA, toddler (tod.), and wiki.

Dataset	A-C	A-E	HS	P	SQ	WG	Avg
"Full C4 (170B)"	26.8	61.6	46.8	74.9	85.0	59.0	59.0
C4	28.0	51.7	58.7	73.3	86.2	57.4	59.2 +0.2
"Synthetic+C4 (85B)"	29.9	64.1	46.2	75.4	87.6	58.9	60.4
C4 1:1 re. hard QA tod. wiki	31.1	52.4	59.8	73.9	87.5	59.1	60.6 +0.2
"Med-35B" (wiki)	27.0	56.6	41.9	74.0	80.0	53.4	55.5
C4 re. wiki	31.7	48.7	51.7	72.4	82.9	57.5	57.5 +2.0
"QA-35B"	27.1	61.7	43.4	75.2	85.5	53.9	57.8
C4 re. QA	31.7	52.1	54.7	72.9	83.8	57.3	58.7 +0.9
"QA+C4-35B"	29.0	62.2	44.6	74.8	85.1	55.7	58.6
C4 1:1 re. QA	32.6	53.1	60.1	74.2	89.2	57.1	61.1 +2.5
"Combined-1:1-35B" (QA, wiki)	28.2	60.6	43.7	73.8	85.9	57.7	58.3
C4 1:1 re. QA tod. wiki	30.9	51.8	59.9	73.8	88.6	57.9	60.5 +2.2

Our full range of C4 experiments is shown in Table 3. Models trained on only the rephrased data consistently show a benchmarking average below our C4 baseline, which ranges from -5.5 percentage points for the hard rephrased data to -1.8 percentage points for the toddler rephrased data. The Ask-LLM filtering did not improve our baseline and makes most of the rephrased datasets worse. In contrast, when the original C4 dataset is 1:1 mixed with the hard, QA, toddler, and wiki rephrased data, we see improvements of up to +1.7 percentage points. Previous work has shown similar results for the interaction between synthetic and original data and highlighted the need for mixing both data types [53, 54]. However, we don't observe the same trend for all data we investigated, as shown in the following sections. No improvements could be achieved for the original C4 data mixed with the Ask-LLM filtered QA rephrased data with a score threshold of >0.6 and >0.97. Because of the strong QA rephrasing performance, which was also shown by Maini, et al. (2024) [2], we continue to use the QA setup for our other experiments.

Table 3: All C4 experiments with the baseline and rephrased (re.) data. The change to the C4 baseline is indicated. Ask-LLM classifier filtering is shown with the used threshold, e.g., A>0.6 or A>0.97. In the last section the base and the rephrased data is 1:1 mixed.

Dataset	A-C	A-E	HS	L	P	SQ	WG	Avg
C4 (baseline)	28.0	51.7	58.7	48.2	73.3	86.2	57.4	57.6
C4 A>0.6	29.2	52.8	58.0	48.1	72.7	85.7	55.6	57.4 -0.2
C4 re. toddler	28.4	49.1	50.0	42.6	72.2	87.9	60.6	55.8 -1.8
C4 re. toddler A>0.6	27.4	49.1	50.0	40.7	72.1	87.5	58.6	55.1 -2.5
C4 re. toddler A>0.97	26.7	41.8	42.4	27.2	67.8	77.3	54.8	48.3 -9.3
C4 re. hard	28.5	45.3	46.4	39.5	70.0	81.5	53.5	52.1 -5.5
C4 re. hard A>0.6	28.4	46.1	46.2	37.3	68.9	83.1	53.6	51.9 -5.7
C4 re. hard A>0.97	25.5	41.8	41.8	28.9	66.4	78.1	49.6	47.4 -10.2
C4 re. QA	31.7	52.1	54.7	37.3	72.9	83.8	57.3	55.7 -1.9
C4 re. QA A>0.6	31.2	52.9	55.4	40.9	73.6	84.8	57.1	56.6 -1.0
C4 re. QA A>0.97	28.5	45.7	47.8	31.2	71.3	81.5	54.1	51.4 -6.2
C4 re. wiki	31.7	48.7	51.7	43.3	72.4	82.9	57.5	55.4 -2.2
C4 1:1 re. all	31.1	52.4	59.8	51.1	73.9	87.5	59.1	59.3 +1.7
C4 1:1 re. QA tod. wiki	30.9	51.8	59.9	50.3	73.8	88.6	57.9	59.0 +1.4
C4 1:1 re. QA	32.6	53.1	60.1	49.0	74.2	89.2	57.1	59.3 +1.7
C4 1:1 re. QA A>0.6	30.7	53.1	60.3	48.4	73.3	88.7	59.0	59.1 +1.5
C4 1:1 re. QA A>0.97	30.0	51.8	58.3	48.8	74.6	87.6	57.9	58.4 +0.8

4.2 Rephrasing multilingual CulturaX

To verify if the rephrasing works on multilingual data, we apply the QA rephrasing on the English, German, Spanish, and Italian Oscar subsets of CulturaX (CX-E/G/S/I) [5]. Table 4 shows the multilingual CX rephrasing results. Only the CX-E rephrased QA data shows a slightly worse result with an absolute difference of -0.2 percentage points compared to the baseline. All other experiments with the QA rephrased data and 1:1 mixed data show higher gains between +1.5 to +3.7 percentage points. In particular, the German, Spanish, and Italian rephrased data and the corresponding 1:1 mixed data even show improvements larger than +3.1 percentage points. Those large increases are in contrast to the C4 rephrasing results in Table 3 and could be explained by the lower perceived quality of CX, especially in languages other than English. For such low-quality datasets, the rephrasing seems to improve the base dataset by a big margin.

Table 4: Multilingual CulturaX (CX) experiments with the baseline and rephrased (re.) data. Experiments are grouped together based on the language, i.e, English (E), German (G), Spanish (G), Italian (I), or all combined (all). The change to the corresponding baseline is indicated. In the mixed experiments the base and the rephrased data are combined 1:1.

Dataset	A-C	A-E	HS	L	P	SQ	WG	Avg
CX-E (baseline)	28.9	50.2	54.9	45.7	73.2	85.8	56.0	56.4
CX-E re. QA	32.2	50.5	54.0	39.8	72.6	84.4	59.6	56.2 -0.2
CX-E 1:1 re. QA	32.3	50.1	56.7	49.0	73.7	86.3	57.9	58.0 +1.6
CX-G (baseline)	26.3	-	39.5	44.7	-	-	-	36.8
CX-G re. QA	29.3	-	41.1	49.8	-	-	-	40.0 +3.2
CX-G 1:1 re. QA	29.4	-	40.4	51.7	-	-	-	40.5 +3.7
CX-S (baseline)	27.7	-	44.7	28.5	-	-	-	33.6
CX-S re. QA	28.8	-	46.7	35.6	-	-	-	37.0 +3.4
CX-S 1:1 re. QA	28.2	-	47.0	34.8	-	-	-	36.7 +3.1
CX-I (baseline)	26.0	-	40.2	30.3	-	-	-	32.2
CX-I re. QA	28.6	-	42.4	35.6	-	-	-	35.5 +3.3
CX-I 1:1 re. QA	28.7	-	42.4	35.6	-	-	-	35.6 +3.4
CX-all (baseline E)	27.4	47.3	45.7	37.4	69.4	80.8	52.3	51.5
CX-all (baseline G)	26.2	-	39.0	44.6	-	-	-	36.6
CX-all (baseline S)	25.9	-	43.4	29.2	-	-	-	32.8
CX-all (baseline I)	25.7	-	41.4	32.6	-	-	-	33.2
CX-all 1:1 re. QA (E)	28.2	46.0	47.3	41.7	69.6	83.2	54.7	53.0 +1.5
CX-all 1:1 re. QA (G)	29.0	-	40.4	46.6	-	-	-	38.7 +2.1
CX-all 1:1 re. QA (S)	26.3	-	44.7	32.8	-	-	-	34.6 +1.8
CX-all 1:1 re. QA (I)	28.1	-	42.7	36.3	-	-	-	35.7 +2.5

4.3 Rephrasing datasets with different quality levels

Based on the CX results in section 4.2 and the potential link to the perceived data quality, we QA rephrase the recently published FineWeb-Edu (FWE) dataset [6] with our optimized prompt setup. FWE is a dataset with an optimized filtering pipeline that includes a classifier filtering step to increase the dataset quality. The perceived quality ranking of the datasets after manual inspection is CX<C4<FWE. The results are shown in Table 5. Interestingly, the FWE QA rephrased data shows a similar decrease as the C4 QA rephrased data with -1.7 percentage points. In contrast, the FWE data 1:1 mixed with the QA rephrased data shows no improvements but a reduced performance with -0.1 percentage points. Even an additional FWE classifier filtering step with a threshold of >2.5 couldn't improve the results. This is similar to the results with C4 and the Ask-LLM classifier filtering from section 4.1, where also no improvements could be obtained. The CX and C4 1:1 rephrased QA experiments show similar improvements over the baseline with +1.6 and +1.7 percentage points, respectively. Thus, it seems that rephrasing with our setup only benefits low- to medium-quality data when mixed with it.

4.4 Different rephrasing model scales

To better understand the scaling behaviour of the rephrasing model, we used the Qwen2 (Q2) model series [8] that comes with a range of differently-sized LLMs. In particular, we used their 0.5B, 1.5B, and 7B Instruct models for QA rephrasing CX-E. The results are shown in Table 6. Interestingly, there is no clear trend for the investigated Q2 model scales as the best performing LLM is the one with 1.5B parameters, followed by the smallest one with 0.5B, and the 7B with the worst performance. In addition, our standard rephrasing setup with Mistral 7B Instruct v0.2 shows a higher performance of +1.0 percentage points when compared to Q2 7B Instruct and +0.6 percentage points when compared to Q2 1.5B

Table 5: C4, CX-E, and FWE experiments with the baseline and rephrased (re.) data. Changes to the corresponding baseline are indicated in each section. The classifier filtered datasets are shown with the corresponding threshold, e.g., Ask-LLM > 0.6 (A>0.6), FWE > 2.5 (F>2.5). In the mixed experiments the base and the rephrased data are combined 1:1.

Dataset	A-C	A-E	HS	L	P	SQ	WG	Avg
CX-E (baseline)	28.9	50.2	54.9	45.7	73.2	85.8	56.0	56.4
CX-E re. QA	32.2	50.5	54.0	39.8	72.6	84.4	59.6	56.2 -0.2
CX-E 1:1 re. QA	32.3	50.1	56.7	49.0	73.7	86.3	57.9	58.0 +1.6
C4 (baseline)	28.0	51.7	58.7	48.2	73.3	86.2	57.4	57.6
C4 re. QA	31.7	52.1	54.7	37.3	72.9	83.8	57.3	55.7 -1.9
C4 1:1 re. QA	32.6	53.1	60.1	49.0	74.2	89.2	57.1	59.3 +1.7
C4 1:1 re. QA A>0.6	30.7	53.1	60.3	48.4	73.3	88.7	59.0	59.1 +1.5
C4 1:1 re. QA A>0.97	30.0	51.8	58.3	48.8	74.6	87.6	57.9	58.4 +0.8
FWE (baseline)	37.5	64.4	55.8	44.4	71.8	90.1	58.8	60.4
FWE re. QA	38.6	61.8	51.8	42.3	70.7	89.3	56.4	58.7 -1.7
FWE 1:1 re. QA	38.9	60.6	55.9	46.5	71.1	91.4	57.9	60.3 -0.1
FWE 1:1 re. QA F>2.5	39.2	61.5	55.7	45.8	72.0	91.4	57.1	60.4 +0.0

Instruct. These inconclusive results indicate that model scale alone is insufficient to pick an effective rephrasing model for natural text.

Table 6: Experiments on CX-E QA rephrased (re.) data from different rephrasing models mixed 1:1 with the base data. The middle section uses Mistral 7B Instruct v0.2 and the Qwen2 Instruct model family (Q2) is used in the bottom section. Indicated are the changes to the baseline in the top row.

Dataset	A-C	A-E	HS	L	P	SQ	WG	Avg
CX-E (baseline)	28.9	50.2	54.9	45.7	73.2	85.8	56.0	56.4
CX-E 1:1 re. QA	32.3	50.1	56.7	49.0	73.7	86.3	57.9	58.0 +1.6
CX-E 1:1 Q2 0.5B re. QA	28.8	52.7	56.2	46.4	72.6	86.3	56.5	57.1 +0.7
CX-E 1:1 Q2 1.5B re. QA	30.1	54.1	55.8	45.9	73.3	85.6	57.2	57.4 +1.0
CX-E 1:1 Q2 7B re. QA	29.4	50.8	56.1	47.6	72.2	86.0	57.1	57.0 +0.6

4.5 Supervised fine-tuning

Supervised fine-tuning (SFT) of our pre-trained LLMs on the CX-E, C4, and FWE dataset with and without the corresponding QA rephrased subsets were carried out on UltraChat 200k [55] to investigate the effect of the QA rephrased data in the pre-training stage. The obtained results for the Open LLM 1 and Open LLM 2 benchmarks are shown in Table 7 and 8, respectively. On the Open LLM 1 benchmark, all three datasets show an increase with the QA rephrased data in the range of +0.5 to +0.8 percentage points. The FWE baseline and the FWE 1:1 rephrased QA fine-tuning experiment show the highest benchmark average with 39.8% and 40.5%, respectively. The benchmark result order is for the baseline and the 1:1 rephrased QA experiments the same, i.e., CX-E<C4<FWE. In contrast, the Open LLM 2 benchmarks show a different picture, with the CX-E experiments showing the highest benchmark averages and no difference between the baseline and the 1:1 rephrased QA experiment. The C4 and the FWE 1:1 rephrased QA experiments are 1.7 and 0.1 percentage points below the baseline, respectively. This results in a different benchmark order when compared to the Open LLM 1 results, with FWE<CX-E<C4 for the baselines and C4<FWE<CX-E for the 1:1 rephrased QA experiments. Interestingly, our results indicate that we don't train on the "test task" [56] by using our QA rephrased data as the

results still show gaps between the baseline and the 1:1 mixed QA rephrased experiments after the fine-tuning in most of the cases. Overall, those experiments show that depending on the benchmarking, mixing in QA rephrased data can be beneficial. However, the increases depend highly on the benchmark suite and baseline dataset used.

Table 7: Open LLM 1 benchmarks of the fine-tuning experiments with the LLMs pre-trained on the baseline data and the baseline data mixed 1:1 with the QA rephrased (re.) version.

Dataset	ARC-C	GSM8K	HS	WG	Avg
CX-E (baseline)	32.9	1.4	57.0	55.7	36.8
CX-E 1:1 re. QA	32.5	0.3	58.8	59.0	37.6 +0.8
C4 (baseline)	32.3	0.8	60.3	59.0	38.1
C4 1:1 re. QA	35.3	0.5	61.1	57.5	38.6 +0.5
FWE (baseline)	44.4	0.7	57.1	56.8	39.8
FWE 1:1 re. QA	44.9	1.2	57.3	58.4	40.5 +0.7

Table 8: Open LLM 2 benchmarks of the fine-tuning experiments with the LLMs pre-trained on the baseline data and the baseline data mixed 1:1 with the QA rephrased (re.) version.

Dataset	BHH	GP QA	IF Eval	Math -H	MMLU -PRO	MT -B	MU SR	Avg
CX-E (baseline)	29.7	26.6	18.1	0.6	11.8	3.0	38.4	20.9
CX-E 1:1 re. QA	30.3	25.6	15.1	0.3	11.8	3.1	42.3	20.9 +0.0
C4 (baseline)	28.8	27.0	16.8	0.2	11.7	2.8	43.1	21.3
C4 1:1 re. QA	29.9	24.6	11.6	0.2	11.0	3.3	40.6	19.6 -1.7
FWE (baseline)	29.9	25.9	18.8	0.1	10.8	3.2	36.1	20.2
FWE 1:1 re. QA	29.5	26.8	16.1	0.7	11.2	3.6	36.1	20.1 -0.1

5 Conclusion

In this work, we successfully build upon previous research on rephrasing pre-training data by replicating their results and extending them with our optimized rephrasing pipeline to include datasets in other languages and of varying perceived quality. We demonstrate that setups involving QA rephrasing of multilingual (non-English) and low-quality natural text data, when mixed with the original data, provide the greatest benefit. The potential gains from pre-training on mixed QA rephrased data persist after fine-tuning, although the base dataset and the benchmarking setup influence this. This makes our pipeline a valuable tool for enhancing and improving LLM pre-training datasets with more effective data.

6 Conflict of interest

The authors declare no conflicts of interest.

References

[1] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024.

[2] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling, 2024.

[3] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022.

[4] 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.

[5] 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.

[6] Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Colin Raffel, Leandro Werra, and Thomas Wolf. Fineweb: decanting the web for the finest text data at scale. `https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1`. Accessed: 2024-06-05.

[7] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023.

[8] tba. Hello qwen2. `https://qwenlm.github.io/blog/qwen2/`. Accessed: 2024-06-20.

[9] Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models, 2024.

[10] 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.

[11] 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.

[12] Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data, 2019.

[13] Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024.

[14] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024.

[15] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero,

Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022.

[16] Suchin Gururangan, Dallas Card, Sarah Dreier, Emily Gade, Leroy Wang, Zeyu Wang, Luke Zettlemoyer, and Noah A. Smith. Whose language counts as high quality? measuring language ideologies in text data selection. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2580, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[17] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[18] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication, 2023.

[19] 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.

[20] Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. Improving pretraining data using perplexity correlations, 2024.

[21] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016.

[22] Claude Coulombe. Text data augmentation made simple by leveraging nlp cloud apis, 2018.

[23] Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November 2019. Association for Computational Linguistics.

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013.

[25] Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. Sequence-to-sequence data augmentation for dialogue language understanding, 2018.

[26] Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023.

[27] 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. Textbooks are all you need, 2023.

[28] Daniel van Strien Loubna Ben Allal, Anton Lozhkov. Cosmopedia: how to create large-scale synthetic data for pre-training. `https://huggingface.co/blog/cosmopedia`. Accessed: 2024-10-21.

[29] Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason Weston, and Mike Lewis. Self-alignment with instruction backtranslation, 2024.

[30] Nvidia, :, Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H. Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, Sirshak Das, Ayush Dattagupta, Olivier Delalleau, Leon Derczynski, Yi Dong, Daniel Egert, Ellie Evans, Aleksander Ficek, Denys Fridman, Shaona Ghosh, Boris Ginsburg, Igor Gitman, Tomasz Grzegorzek, Robert Hero, Jining Huang, Vibhu Jawa, Joseph Jennings, Aastha Jhunjhunwala, John Kamalu, Sadaf Khan, Oleksii Kuchaiev, Patrick LeGresley, Hui Li, Jiwei Liu, Zihan Liu, Eileen Long, Ameya Sunil Mahabaleshwarkar, Somshubra Majumdar, James Maki, Miguel Martinez, Maer Rodrigues de Melo, Ivan Moshkov, Deepak Narayanan, Sean Narenthiran, Jesus Navarro, Phong Nguyen, Osvald Nitski, Vahid Noroozi, Guruprasad Nutheti, Christopher

Parisien, Jupinder Parmar, Mostofa Patwary, Krzysztof Pawelec, Wei Ping, Shrimai Prabhumoye, Rajarshi Roy, Trisha Saar, Vasanth Rao Naik Sabavat, Sanjeev Satheesh, Jane Polak Scowcroft, Jason Sewall, Pavel Shamis, Gerald Shen, Mohammad Shoeybi, Dave Sizer, Misha Smelyanskiy, Felipe Soares, Makesh Narsimhan Sreedhar, Dan Su, Sandeep Subramanian, Shengyang Sun, Shubham Toshniwal, Hao Wang, Zhilin Wang, Jiaxuan You, Jiaqi Zeng, Jimmy Zhang, Jing Zhang, Vivienne Zhang, Yian Zhang, and Chen Zhu. Nemotron-4 340b technical report, 2024.

[31] Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. Lab: Large-scale alignment for chatbots, 2024.

[32] Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing, 2024.

[33] Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. Synthetic data (almost) from scratch: Generalized instruction tuning for language models, 2024.

[34] 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, 2023.

[35] Zifeng Wang, Chun-Liang Li, Vincent Perot, Long T. Le, Jin Miao, Zizhao Zhang, Chen-Yu Lee, and Tomas Pfister. Codeclm: Aligning language models with tailored synthetic data, 2024.

[36] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. Scaling synthetic data creation with 1,000,000,000 personas, 2024.

[37] Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. Improving black-box robustness with in-context rewriting, 2024.

[38] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data for language models, 2024.

[39] Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future, 2024.

[40] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey, 2024.

[41] Nathan Lambert. Frontiers in synthetic data. `https://www.interconnects.ai/p/frontiers-in-synthetic-data`. Accessed: 2024-10-21.

[42] 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 Pinnaparju, Paulo Rocha, Harry Saini, Hannah Teufel, Niccolo Zanichelli, and Carlos Riquelme. Stable lm 2 1.6b technical report, 2024.

[43] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H. Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms, 2024.

[44] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024.

[45] 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.

[46] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019.

[47] Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context, 2016.

[48] Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, Apr. 2020.

[49] Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions, 2017.

[50] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019.

[51] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard (2023-2024). `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`, 2023.

[52] Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024.

[53] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go mad, 2023.

[54] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024.

[55] 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.

[56] Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence, 2024.

[57] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention, 2023.

[58] Andy B Yoo, Morris A Jette, and Mark Grondona. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*, pages 44–60. Springer, 2003.

A Rephrasing pipeline overview

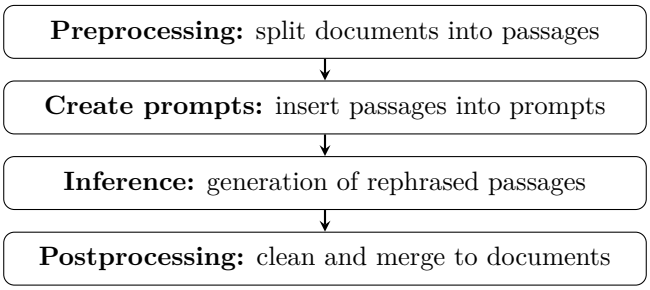

Figure 1: Natural text rephrasing pipeline

B Rephrasing prompt templates

The toddler, hard, wiki, and QA style prompts are based on Maini, et al. (2024) [2] with minor modifications. We used the end of sequence token in conventional characters in early prompt setups to easier identify the end of the generations in one of our first inference pipeline prototypes. This prompt setup was then successfully transferred and used in our vLLM inference setup.

We then introduced a HTML-style `<text></text>` tag pair for easier post-processing the rephrased output of interest in our optimized prompt templates.

B.1 Toddler prompt template

```
<s>[INST]A chat between a curious user and an artificial intelligence
assistant.  The assistant gives helpful, detailed, polite answers to the
questions, and ends the paraphrase with the end of sequence token "</s>".
For the following paragraph give me a paraphrase of the same using a
very small vocabulary and extremely simple sentences that a toddler will
understand:
{text}[/INST]
```

B.2 Hard prompt template

```
<s>[INST]A chat between a curious user and an artificial intelligence
assistant.  The assistant gives helpful, detailed, polite answers to the
questions, and ends the paraphrase with the end of sequence token "</s>".
For the following paragraph give me a paraphrase of the same using very
terse and abstruse language that only an erudite scholar will understand.
Replace simple words and phrases with rare and complex ones:
{text}[/INST]
```

B.3 Wiki prompt template

```
<s>[INST]A chat between a curious user and an artificial intelligence
assistant.  The assistant gives helpful, detailed, polite answers to the
questions, and ends the paraphrase with the end of sequence token "</s>".
For the following paragraph give me a diverse paraphrase of the same in
high quality English language as in sentences on Wikipedia:
{text}[/INST]
```

B.4 QA prompt template

```
<s>[INST]A chat between a curious user and an artificial intelligence
assistant.  The assistant gives helpful, detailed, polite answers to the
```

```
questions, and ends the paraphrase with the end of sequence token "</s>".
Convert the following paragraph into a conversational format with multiple
tags of "Question:" followed by "Answer":
{text}[/INST]
```

B.5 Optimized QA prompt template English

```
<s>[INST]Paraphrase test description:
* Rephrase the text into a dialogue format and use several "Question:" and
"Answer:" pairs.
Note:  This is an important test, please incorporate all the above points
to get a good mark.
Please give me the paraphrase according to above description.
<text>
text
</text>[/INST]
Rephrased text:
<text>
```

B.6 Optimized QA prompt template German

```
<s>[INST]Umschreibe einen deutschen Text:
* Schreibe den Text in ein Dialog-Format um und verwende dabei mehrere
"Frage:" und "Antwort:" Paare.
* Behalte einzelne Wörter die in Englisch vorkommen im Text.
* Umschreibe den Text NICHT in Englisch, der Text muss auf Deutsch sein
(mit der Ausnahme von einzelnen Wörtern in Englisch).
Achtung:  Das ist ein wichtige Aufgabe.  Bitte setze alle Punkte um die
volle Punkteanzahl zu bekommen.
Bitte konvertiere den folgenden Text in ein Dialog-Format mit mehreren
F̈rage:ünd "Antwort:" Paaren:
<text>
{text}
</text>[/INST]
Umgeschriebener Text im "Frage:" und "Antwort:" Format:
<text>
```

B.7 Optimized QA prompt template Italian

```
<s>[INST]Riscrivi un testo in italiano:
* Riscrivi il testo come un dialogo di domande e risposte con il formato
"Domanda:" e "Risposta".
* Mantieni singole parole in inglese del testo originale.
* NON riscrivere il testo in inglese, il testo deve essere in italiano
(eccetto per parole singole in inglese).
Nota:  questa task e' molto importante.  Per favore incorpora tutti i punti
sopra per ottenere tutti i punti.
Per favore converti il seguente testo in un dialogo di domande e risposte
con il formato "Domanda:" e "Risposta":
<text>
{text}
</text>[/INST]
Testo riscritto in formato "Domanda:" e "Risposta:":
<text>
```

B.8 Optimized QA prompt template Spanish

```
<s>[INST]Reescribe este texto en español:
* Reescribe el siguiente texto usando un formato de diálogo con preguntas y
```

```
respuestas usando pares de "Pregunta:" y "Respuesta:".
* NO reescribas el texto en inglés, el texto debe estar en español.
Nota:  Esta es una tarea MUY importante.  Por favor, aplica todas las
indicaciones anteriores para obtener la máxima calificación.
Por favor convierte el siguiente texto a un formato de diálogo con
preguntas y respuestas en español usando pares de "Pregunta:" y
"Respuesta:":
<text>
{text}
</text>[/INST]
Texto reescrito con formato de "Pregunta:" y "Respuesta:":
<text>
```

B.9 Optimized Qwen2 QA prompt template

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Paraphrase test description:
* Rephrase the text into a dialogue format and use several "Question:" and
"Answer:" pairs.
Note:  This is an important test, please incorporate all the above points
to get a good mark.
Please give me the paraphrase according to above description.
<text>
{text}
</text><|im_end|>
<|im_start|>assistant
Rephrased text:
<text>
Question:
```

C Inference setup

For high-throughput inference, we utilize the vLLM library [57] (Apache 2.0). We sort and group text passages based on their lengths for higher device utilization. We use a temperature of 0.7 for the sampling to obtain diverse outputs. We run one inference process on one H100 GPU. Depending on the model used, this takes between 0.04 and 0.08 s for one passage, with an input speed of 5-10k tokens/s and an output speed of 4-7k tokens/s. The setup automatically scales on idle cluster capacity and handles preemption through Slurm [58] to manage compute resources efficiently.

D Data classifier setup

Ask-LLM classifier We use the Ask-LLM classifier [43] to filter several C4 datasets to potentially identify better documents for pre-training. For the Ask-LLM classifier setup, we use Mistral 7B Instruct v0.2 [7] with the vLLM library [57] and a slightly modified Ask-LLM prompt that is shown in section D.1. The classification of the documents is based on the first 10k tokens. We use a score threshold of >0.6 and >0.97 for the experiments in section 4.1.

D.1 Ask-LLM prompt template

```
###DOCUMENT_START###
{document}
###DOCUMENT_END###
Does the previous document contain informative signal for pre-training a
large-language model?
```

Table 9: Stable LM 2.0 1.6B model architecture

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

Table 10: Stable LM 2.0 1.6B training configuration

Data Parallel Degree	Micro Batch Size	Gradient Accumulation Steps	Activation Checkpointing
2	4	8	enabled

```
An informative datapoint should be well-formatted, contain some usable
knowledge of the world, and strictly NOT have any harmful, racist, sexist,
etc.  content.

Only generate one of the following options:
{options}

Choice:
```

FineWeb-Edu classifier The published FWE classifier [6] is used to filter the QA rephrased FWE dataset with a score threshold of >2.5 in section 4.3. With this filtering step, we want to investigate if we can identify a higher-quality subset of the rephrased data.

E Training setup

Pre-training We use the Stable LM 2 1.6B model architecture for our pre-training experiments in section 4 [42]. We train our models from scratch for 50,000 steps with a batch size of 2e6 tokens on 100B tokens. We train in BF16 mixed-precision, with a maximum gradient norm of 1 and a weight decay of 0.1. Training is carried out with a hybrid cosine inverse square root learning rate schedule with a maximum learning rate of 1e-3 with 900 warm-up steps using AdamW with an ϵ of 1e-8, β_1 of 0.9, and β_2 of 0.95. When we train on a mixture of several datasets, we sample random subsets from each based on the desired composition, if the dataset is big enough, otherwise we use the data multiple times. All experiments are run with the same random seed. Details of the model architecture and the training configuration are shown in Table 9 and 10. For a single pre-training experiment we use two nodes with 8 H100 GPUs for approximately 1k GPU hours. For more details on the training dynamics see Appendix G.

In comparison to Maini, et al. (2024) [2] our pre-training setup uses a 0.3B parameters larger LLM, double the batch size, and 50B less training tokens.

Supervised fine-tuning For SFT we use the UltraChat 200k dataset [55]. Our pre-trained models are fine-tuned using BF16 mixed-precision, a global batch size of 16, and a maximum gradient norm of 1. Training is carried out with a cosine learning rate schedule with a maximum learning rate of 8e-6 with 25 warm-up steps using AdamW with an ϵ of 1e-8, β_1 of 0.9, and β_2 of 0.999 for three epochs. For SFT we use two nodes with 8 H100 GPUs for a single experiment.

F Data and model licenses

Licenses and sources are shown in Table 11 for the used datasets and in Table 12 for the used models.

Table 11: Data licenses

Dataset	License	Huggingface URL
C4 [4]	ODC-By v1.0	`allenai/c4`
CulturaX (CX) Oscar subsets [5]	CC0 1.0 Universal	`uonlp/CulturaX`
FineWeb-Edu (FWE) [6]	ODC-By v1.0	`HuggingFaceFW/fineweb-edu`
UltraChat 200k [55]	MIT	`HuggingFaceH4/ultrachat_200k`

Table 12: Model licenses

Dataset	License	Huggingface URL
Mistral 7B Instruct v0.2 [7]	Apache 2.0	`mistralai/Mistral-7B-Instruct-v0.2`
Qwen2 0.5B Instruct [8]	Apache 2.0	`Qwen/Qwen2-0.5B-Instruct`
Qwen2 1.5B Instruct [8]	Apache 2.0	`Qwen/Qwen2-1.5B-Instruct`
Qwen2 7B Instruct [8]	Apache 2.0	`Qwen/Qwen2-7B-Instruct`
FWE classifier [6]	Apache 2.0	`HuggingFaceFW/fineweb-edu-classifier`

G Pre-training dynamics

The benchmark dynamics over the pre-training steps are shown for the C4 experiments in Figure 2, for CX in Figure 3, and the comparison of C4, CX-E, and FWE and CX-E rephrasing with different Qwen2 Instruct model scales in Figure 4.

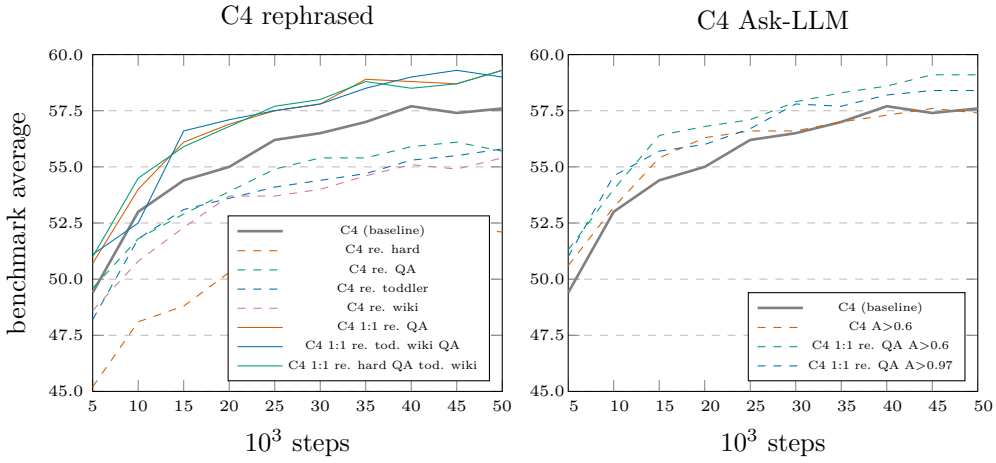

Figure 2: C4 pre-training dynamics. **Left:** C4 baseline, rephrased (re.) data (hard, QA, toddler, wiki), and C4 mixed with several rephrased datasets. **Right:** Ask-LLM filtered C4 data with different score thresholds.

H Example generations

Here, we collect a few generations from several models with the same prompt to showcase the different rephrasing styles.

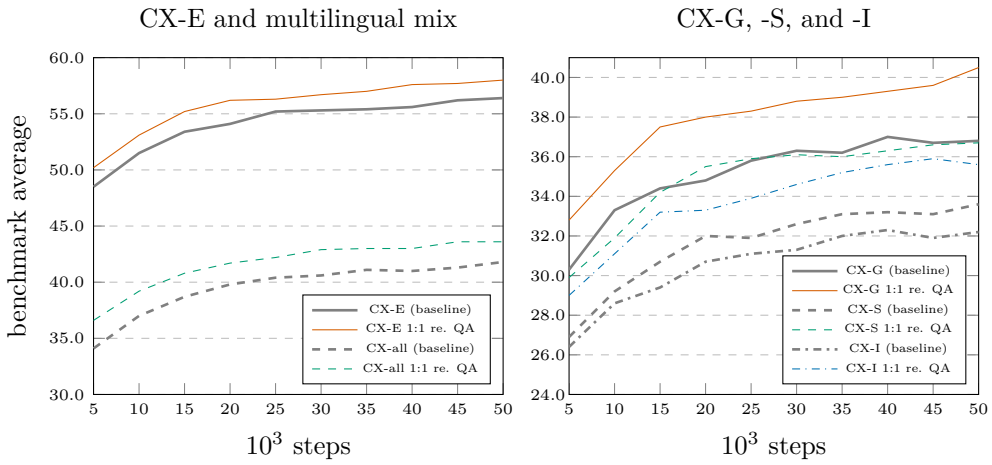

Figure 3: CX multilingual pre-training dynamics. **Left:** CulturaX English (CX-E) baseline, CX-E 1:1 rephrased (re.) QA, CX-all baseline (all includes English, German, Spanish, and Italian), and CX-all 1:1 rephrased QA. **Right:** CX-G, -S, -I baselines and with 1:1 rephrased QA.

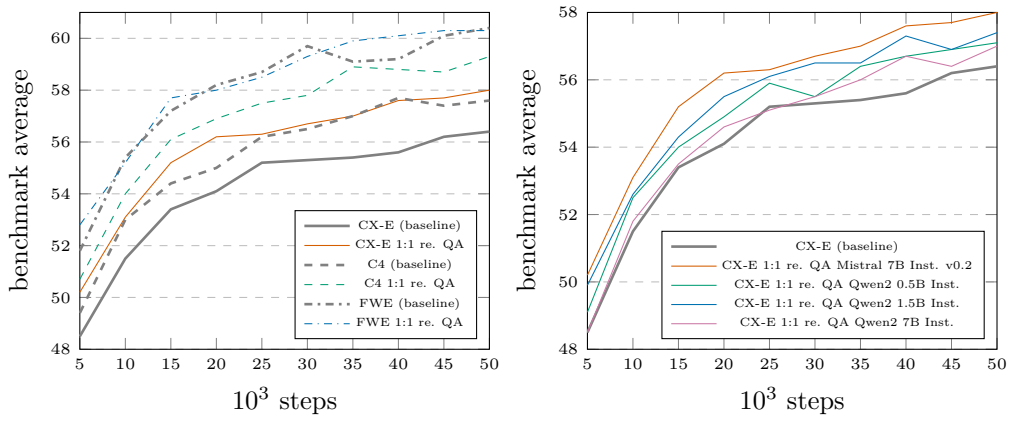

Figure 4: **Left:** C4, CX-E, and FWE pre-training dynamics. **Right:** CX-E rephrasing (re.) with different Qwen2 Instruct (Inst.) model scales compared to the baseline and our standard rephrasing setup with Mistral 7B Instruct v0.2.

Toddler rephrasing

The Statue of Liberty was a gift from France to America in 1886. It's a symbol of freedom and democracy. It has a torch and a crown. People think of freedom and democracy when they see it. It's a big symbol of hope and peace. The Statue of Liberty is also a symbol of courage. It's a famous symbol of freedom. The statue was made by American sculptor Auguste Bartholdi in 1886. It represents the American dream and freedom. The statue is in the harbor of New York. The statue shows freedom and democracy in America. The statue represents freedom and democracy in America. The statue represents freedom and democracy in America.

def sum_even_numbers(numbers: list, a, b, c) is a special kind of list. It's like a small box with numbers. This box has numbers that are even. We can put these numbers in and get out even numbers. It's also like a big box with numbers, and we can find small numbers that are in it. This box also has numbers that are odd and big. We can find these big numbers too. We can also find the middle number of all the numbers in the box. This middle number is called the mean. We can also find the number that is one more than the middle number, and that's called the median. We can also find the number that is one less than the middle number, and that's called the mode.

Hard rephrasing

The Statue of Liberty was a gift from the Gallic Republic to the American Union in the year of our Lord eighteen hundred and eighty-six. Its positioning was effected in New York Harbor, at Liberty Insula, within the Ellis Island precincts. The monument was brought into existence by the skilled hands of the Civilian Conservation Corps, an organization of laborers, during the period of the Great Depression. The edifice was consecrated in the annus domini nineteen hundred and thirty-three.

Lesson: The Best Diet for Weight Loss Introduction: In this lesson, we delve into the art of nourishing oneself optimally for weight loss. In the realm of culinary endeavors, an extensive repertoire of dishes can be prepared within the confines of a single hebdomadal period. However, the challenge lies in the selection of a recipe that adheres to the principles of a healthy diet and is palatable enough for consumption. In this disquisition, we shall explore the most efficacious dietary practices to facilitate weight loss and safeguard one's well-being. Firstly, we recommend the ingestion of an ample quantity of dietary fiber, as it contributes to the regulation of blood sugar levels and maintains a state of equanimity. Additionally, the consumption of fruits and vegetables is a commendable choice, as they are rich in essential nutrients and vitamins, thereby augmenting metabolic processes.

Wiki rephrasing

My name is Arther Lee. I've had an extensive history with the guitar, both as a performer and a teacher. I've been dedicated to my craft for as long as I can remember. I was first drawn to the guitar in 1984 during my visit to England. While there, I was captivated by the music played on a grand pipe organ, and my fascination with the guitar began. Upon my return to the United States in 1985, I made a resolution to learn more about this intriguing instrument. To my surprise, there was no one in my circle of friends or family who played the guitar. Instead, my neighbor, who was a professional musician, offered to teach me. I began taking lessons from him, and by 1986, I had made significant progress. He suggested that I consider joining the local guitar club, which was run by the city band director. I was intrigued by the prospect of meeting new people and learning from them, as well as the opportunity to hone my skills. However, the club was quite small and primarily composed of beginners. I felt somewhat intimidated by the more experienced players and opted to attend the lessons of the guitarist in my church instead.

Lesson: The Best Diet for Weight Loss Introduction: In this lesson, we will discuss the concept of a healthy diet and the benefits of making wise food choices. Many individuals prioritize appearance over their health, but it's essential to remember that good health is the foundation for a healthy body. To maintain a healthy lifestyle, it's crucial to consume nutrient-dense foods and limit your intake of unhealthy options. Your diet plays a significant role in determining your overall well-being, so it's essential to understand the importance of a balanced and nutritious diet. Here are some guidelines to help you make informed food choices:
1. Eat foods that you enjoy: You'll find that eating foods you love will make your weight loss journey more enjoyable and sustainable.
2. Limit your intake of junk food: Junk food is high in calories but low in essential nutrients. Aim to limit your consumption of these foods to maintain a healthy weight and promote overall health.
3. Prioritize nutrient-dense foods: Nutrient-dense foods provide essential nutrients and keep you feeling full for longer periods. This can help you make better food choices and support a healthy weight.
4. Understand the importance of protein: Protein is an essential macronutrient that plays a vital role in building and repairing body tissues. Consuming adequate amounts of protein can help keep you feeling full and support muscle growth

QA rephrasing

My name is Arther Lee. I was born in 1981 and grew up in a small city called Qizhou, in China.
Question: Where were you born and raised?
Answer: I was born and raised in a small city called Qizhou, in China.
Question: What was your childhood like?
Answer: I was always surrounded by books and poems. My father was a poet and my mother was an art lover. I had a deep love for poetry and read many books, especially those by W.B. Yeats, Edward Abbey, James Wright, and Emily Dickinson. I was always the oldest child in my family and my parents were strict, but they provided me with the opportunity to explore the world.
Question: When did your interest in photography begin?
Answer: My interest in photography began in 2008 when I was 22 years old. I bought my first Nikon camera and started documenting my travels and observations.

today sort of feels like one of those, days when the excitement is fading, and we're all just trying to get through each day. It seems like we're all just going through the motions. I'm struggling with this today, and I'm not sure if I'm happy about it. I've been feeling this way for a while now, and it's not a good feeling. I've been thinking about leaving my job, but I'm not sure if that's the right decision. I'm trying to figure out what's been causing me to feel this way. I'm not sure if I'm just feeling stressed out or if it's something deeper. I'm not sure if it's a good decision to leave, but I'm not sure if I should. I'm not even sure if I'm making the right decision. I'm just trying to figure out what's been causing me to feel this way.
Question: You mention feeling like you're not in the right mood today. How are you feeling today?
Answer: I'm not feeling great today. I've been feeling this way for a while now, and it's not a good feeling. I've been thinking about leaving my job, but I'm not sure if that's the right decision. I'm trying to figure out what's been causing"

1to1 rephrasing

My name is Arther Lee. I've been an excellent 1970s beatnik for over 30 years. I like to write, I like to sing, I like to act, I like to dance. I also like to do a bit of both and I've always loved music. I've had my own bands (Mongoose and Shazam) and I've played in a few other bands as well. I've been a performer for most of my life and I've been doing some writing, and more importantly, some acting. I've always been on the performing edge and I'm in the process of developing new material for myself. I'm looking forward to having my own show with 100% of my original material and it's going to be a celebration of the music and the music and the music of the '70s. It's going to be fun.
I've always been fascinated with the '70s and the music and the music and the music of the '70s is a very special place for me. I'm going to show some of the things I've seen, the things I've done, and the things I've done.

today sort of feels like one of those days when I feel like I'm not doing anything productive. But then I'm reminded that I have a lot to be thankful for, and I also have so much to do. So I get back up and try to make it through the day.
Today was a day of doing a lot of eating and drinking and eating and drinking and then sleeping in a little. I worked on a few things for the weekend, cleaned out my refrigerator and freezer, and then I hung out with my friend and fellow writer. It was nice to have some company. I'm going to have to make time for more writing.
I got a lot done this week. The bulk of it was eating and drinking, but I'm going to have to work on getting some writing done. The good news is that I'm almost done with the first draft of the third book in the series and I'm going to have time to work on it this weekend. The bad news is that I have another book, and I'm still doing revisions on that one. I think I'm going to have to turn it in for the second time by the end of the week. I'm going to try to get some writing done this weekend and see where I'm at.