DecorateLM: Data Engineering through Corpus Rating, Tagging, and Editing with Language Models

Ranchi Zhao[1]*, Zhen Leng Thai[2]*, Yifan Zhang[1]*, Shengding Hu[2]*,
Yunqi Ba[1], Jie Zhou[1], Jie Cai[1], Zhiyuan Liu[2]†, Maosong Sun[2]†,
[1]Modelbest Inc, [2]Department of Computer Science and Technology, Tsinghua University
{ranchizhao,thaizhenleng123,yifanzhang634,shengdinghu}@gmail.com

Abstract

The performance of Large Language Models (LLMs) is substantially influenced by the pretraining corpus, which consists of vast quantities of unsupervised data processed by the models. Despite its critical role in model performance, ensuring the quality of this data is challenging due to its sheer volume and the absence of sample-level quality annotations and enhancements. In this paper, we introduce *DecorateLM*, a data engineering method designed to refine the pretraining corpus through data rating, tagging and editing. Specifically, *DecorateLM* rates texts against quality criteria, tags texts with hierarchical labels, and edits texts into a more formalized format. Due to the massive size of the pretraining corpus, adopting an LLM for decorating the entire corpus is less efficient. Therefore, to balance performance with efficiency, we curate a meticulously annotated training corpus for *DecorateLM* using a large language model and distill data engineering expertise into a compact 1.2 billion parameter small language model (SLM). We then apply *DecorateLM* to enhance 100 billion tokens of the training corpus, selecting 45 billion tokens that exemplify high quality and diversity for the further training of another 1.2 billion parameter LLM. Our results demonstrate that employing such high-quality data can significantly boost model performance, showcasing a powerful approach to enhance the quality of the pretraining corpus.

1 Introduction

The advent of Large Language Models (LLMs) has ushered in transformative changes across various domains of artificial intelligence (Brown et al., 2020; Chowdhery et al., 2023), from natural language processing to complex task execution (Qian et al., 2023). The backbone of these models' effectiveness lies in their training processes, specifically in the quality and composition of their pre-training corpora (Penedo et al., 2023; Le Scao et al., 2023). Traditionally, LLMs are pre-trained on vast datasets composed of billions of tokens harvested from diverse text sources.

Data quality is of vital importance for training LLM (Zhou et al., 2024). However, acquiring high-quality data is a formidable challenge due to the sheer volume and unstructured nature of it.

The reliance on large-scale unsupervised data leads to the inclusion of numerous low-quality texts within the training data. This infusion of poor-quality data can adversely affect the models' learning processes, resulting in performance deficiencies and limitations in their applicability. However, the existing methods for curating and enhancing the quality of such datasets are often inadequate. They typically lack the capacity to scale to the size required while maintaining or improving data quality, primarily due to the absence of fine-grained annotations and the impracticality of manual oversight.

Addressing these challenges requires innovative approaches that can scale with the data requirements of LLMs while ensuring enhancements in data quality. This paper introduces *DecorateLM*, a comprehensive data engineering methodology designed to refine the pretraining corpus through a systematic "decorating" process. The term "decorating" in this context refers to a series of processes aimed at enriching the data with additional metadata, improving its structure, and ensuring its relevance and quality.

DecorateLM employs a three-phase strategy to accomplish these goals. The first phase, rating, involves evaluating texts against a predefined set of quality criteria. These criteria are designed to assess the educational value, expertise, fact and trivia, reasoning level, scarcity, structural format, story-likeness and subjectivity of texts. The second

---

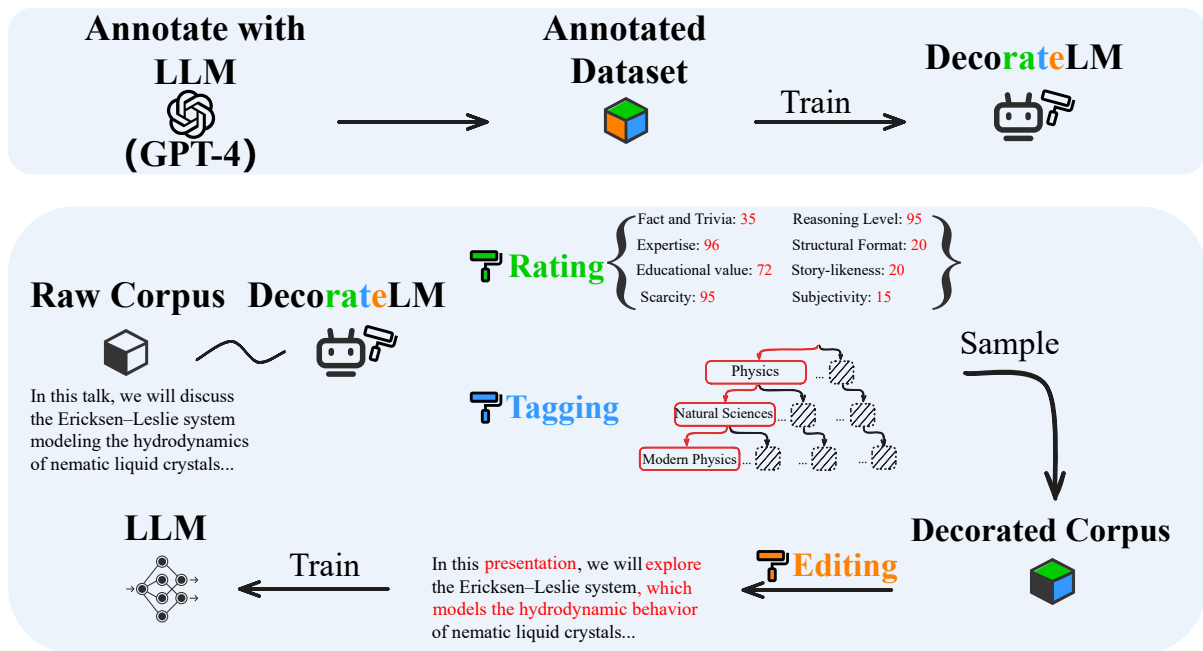*Equal contribution.
†Corresponding author.

Figure 1: We utilize GPT-4 to assemble an annotated training corpus and integrate data engineering expertise into DecorateLM. DecorateLM is then used to process 100 billion tokens from the raw corpus, sampling 45 billion tokens using its rating and tagging capabilities to create what we refer to as the Decorated corpus. We further enhance the Decorated corpus by applying DecorateLM's editing features, making it more suitable for LLM training.

phase, tagging, categorizes the texts using a hierarchical label system that reflects the content of the data. This labeling enhances data management and retrieval efficiency, a key aspect of iterative training processes. The final phase, editing, involves revising and standardizing texts to meet higher linguistic standards of formality and clarity.

To implement this methodology effectively, we curate a specialized training corpus using pretrained LLMs to preprocess and initially rate potential data samples. This approach leverages the model's capabilities to perform initial assessments at scale. We then distill our data engineering expertise into a small language model (SLM)—which is optimized for more detailed and nuanced data processing tasks. We name this SLM as the *DecorateLM*. Using *DecorateLM*, we enhance 100 billion tokens from our initial datasets, selecting 45 billion tokens that exhibit optimal quality and diversity. These tokens are subsequently used to train LM to demonstrate DecorateLM's effectiveness.

The results from our study underscore the substantial benefits of using high-quality, well-curated data in training LLMs. Not only do these results demonstrate improved model performance, but they also suggest that *DecorateLM* offers a scalable and effective solution to one of the most pressing issues in modern AI—enhancing the quality of training

datasets amid expanding data requirements.

2 Related Work

In recent years, the quality and selection of data for training language models receive considerable attention. Researchers propose various methodologies to assess, select, and improve high-quality data, with the goal of enhancing both the performance and efficiency of models (Elazar et al., 2023; Longpre et al., 2023; Xie et al., 2023; Li et al., 2024).

Data Annotation and Rating. QuRating, DEITA, and ALPAGASUS are employed for data annotation, each utilizing distinct methodologies to enhance training via refined rating scores (Wettig et al., 2024; Liu et al., 2023; Chen et al., 2023). Phi-1 and MoDS use GPT-4 and DeBERTa to improve educational data and precise data selection, accelerating learning and fine-tuning (Gunasekar et al., 2023; Du et al., 2023).

Domain Diversity in Data. INSTAG introduces a detailed tagging system for diverse SFT data, improving MT-Bench scores with less data (Lu et al., 2023). Phi-1.5 extends Phi-1 by adding synthetic data across multiple domains in a textbook style (Li et al., 2023b).

Data Optimization for Model Training. Studies show that models can perform well with smaller datasets and less computing. WRAP maintains per-
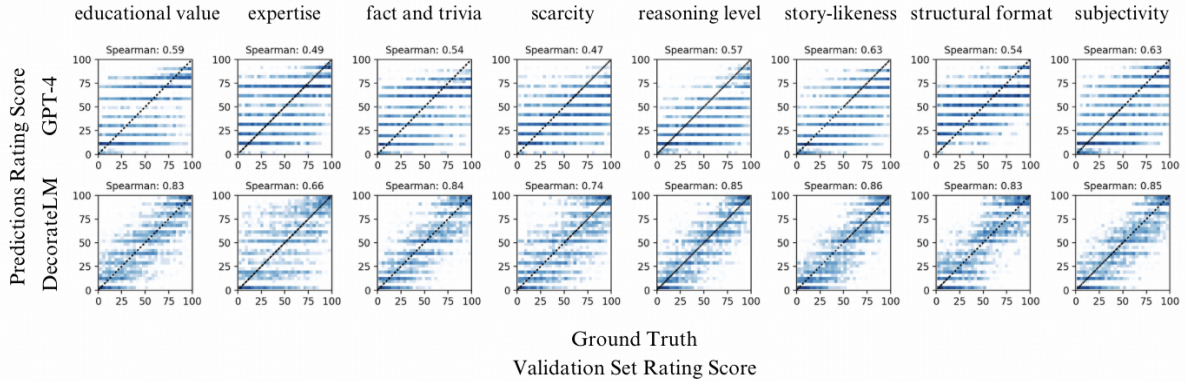
Figure 2: The Spearman correlations between model ratings and ground truth of validation set. Specifically, the x-axis represents the ground truth rating scores of the data. The y-axis represents the prediction rating scores of GPT-4 and DecorateLM after evaluating the validation set. Rating scores generated by GPT-4 are more discrete and inaccurate compared to DecorateLM.
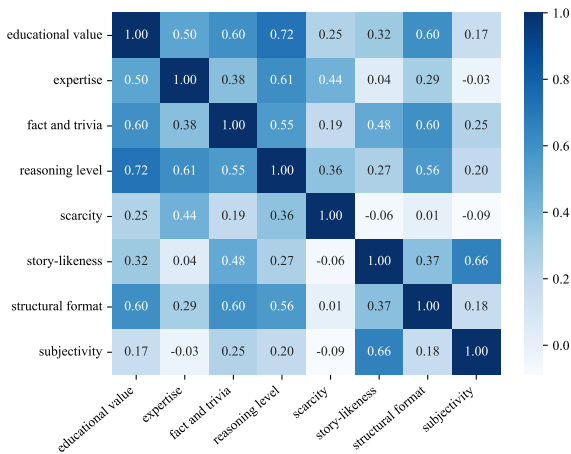

Figure 3: Spearman correlation coefficients between various rating criteria. The correlations align with intuitive expectations. For instance, data with higher educational value often exhibits enhanced reasoning levels, which, in turn, enhances their comprehensibility.

formance with fewer resources on the C4 dataset, and TinyStories uses simple vocabulary for quicker learning (Maini et al., 2024; Eldan and Li, 2023). Additionally, Phi-3 uses a two-stage training with web and synthetic data to improve reasoning and specialized skills (Abdin et al., 2024).

3 Method

3.1 Framework

In this section, we detail the methodology of DecorateLM, which is designed for sample-level annotation and enhancement. The framework of DecorateLM consists of three distinct phases: rating, tagging, and editing. During the rating phase, DecorateLM assigns numeric scores to a text based on predefined quality dimensions. In the tagging phase, DecorateLM predicts hierarchical tags at three levels for the text. In the editing phase, DecorateLM rephrases the text to present alternative narratives, thereby facilitating the model's acquisition of core knowledge from varied perspectives.

The training pipeline of DecorateLM incorporates both a teacher model and a student model. The teacher model, which is larger, excels in processing detailed instructions related to text quality. However, its slower processing speed limits its practicality for annotating or editing extensive pretraining corpora. To address this, knowledge from the teacher model is distilled into a more compact student model to enhance efficiency. Distinct distillation strategies are employed for each of the three phases. The rating and tagging phases, which involve processing the entire raw corpus and generating concise annotations, exhibit similar input-output dynamics. Consequently, DecorateLM is configured to manage these two phases concurrently to optimize efficiency, instead of leveraging two separate models. For the editing phase, a separate distillation process is implemented to distill the knowledge required for effective rephrasing into another model of DecorateLM.

3.2 Rating

High-quality training data is crucial for developing powerful language models. However, the ideal properties that constitute an optimal training corpus remain challenging to characterize comprehensively. To achieve robust language understanding and generation capabilities, language models should be trained on high-quality data meticulously

Figure 4: Word cloud of tags. The size of each tag is proportional to its frequency in the annotated dataset. Tags are color-coded based on their levels: first-level tags in dark blue, second-level tags in medium blue, and third-level tags in light blue.

Model	First	Second	Third
DecorateLM	92.1	75.6	62.3
GPT-4	93.6	77.3	68.5

Table 1: Comparison of tagging accuracy between DecorateLM and GPT-4 across three hierarchical levels on the validation set. GPT-4, lacking prior knowledge of the designed tagging hierarchy, is provided with the relevant labels for each level through prompts in successive rounds of interaction.

curated based on diverse criteria that capture the essential and abstract qualities of natural language texts.

Criteria. To assess the quality of texts, we define eight evaluative criteria that quantitatively measure the contributions of a text to model training from multiple perspectives. For each criterion, data samples are assigned a quantitative score, enabling an objective evaluation across the various criteria.

1. *Educational Value* evaluates whether the content is suitable for educational purposes, specifically its utility in textbooks. It assesses the clarity, detail, and comprehensibility of explanations and guiding principles.

2. *Expertise* measures the depth of knowledge that content reflects, typically possessed by subject matter experts.

3. *Fact&Trivia* focuses on the accuracy of factual information presented in the content, which does not necessarily require specialized expertise to understand.

4. *Reasoning Level* assesses the necessity for high-level reasoning, sequential thought pro-

cesses, or chain of thought (Wei et al., 2022) capabilities in the content.

5. *Scarcity* targets accurate yet relatively unknown information that is typically familiar only to a select few due to its specialized, niche, or obscure nature.

6. *Structural Format* evaluates the organization and structure of data, such as the use of numbered lists, bulleted lists, and markdown formatting.

7. *Story-likeness* assesses whether the content narrates a story or describes a scenario.

8. *Subjectivity* focuses on content with personal opinions and conversations.

Annotated Dataset Construction. In alignment with the established criteria, we annotate a set of carefully selected samples using GPT-4 to form the annotated dataset. Considering the inaccuracy of LLMs in assigning precise quality scores (Zheng et al., 2024), we adopt a pairwise comparison method. Inspired by QuRating (Wettig et al., 2024), this work employs the Bradley-Terry (B-T) model (Bradley and Terry, 1952) to derive preference probabilities from pairwise comparisons. All prompts used in the rating phase are displayed in Appendix A.1. Subsequently, we normalize these probabilities by sorting them and applying a linear transformation to map them onto a uniform rating scale from 0 to 100, thereby establishing the final scores for each criterion.

Analysis. Upon acquiring the meticulously curated annotated dataset, we proceed to train DecorateLM, with the training details provided in Appendix B.1. A validation set is segregated prior to training. DecorateLM is employed to assign scores to each data sample. For a fair comparison, we also use GPT-4 to assign numeric scores to these samples. Then we compute the Spearman correlation coefficient between the model-provided scores and the ground truth annotation from the B-T model. As depicted in Figure 2, GPT-4, untrained for the rating task, demonstrates *inferior* scoring performance compared to DecorateLM.

In the analysis presented in Figure 3, we compute the Spearman correlation coefficients between various rating criteria. The results reveal a modest positive correlation across most pairs of criteria,
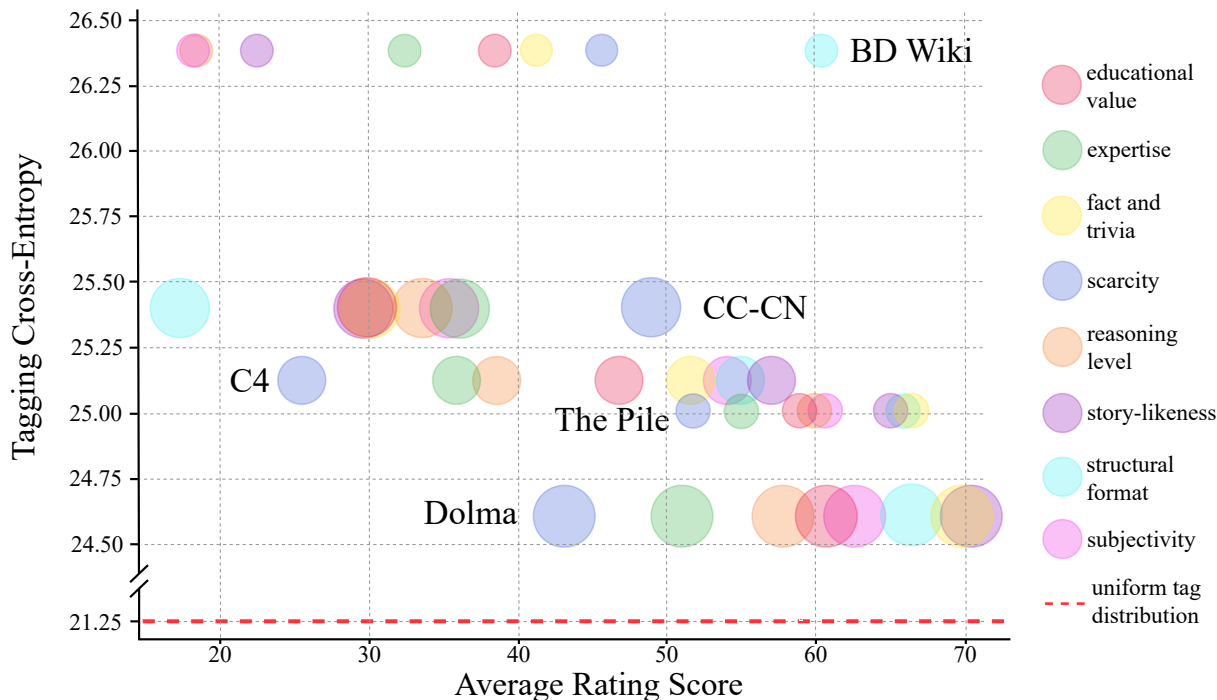
Figure 5: Evaluation of dataset rating and tagging quality using DecorateLM. The x-axis denotes the average rating of each dataset across specified dimensions, whereas the y-axis represents the cross-entropy of tags from predefined tagging system. The circle size correlates with the dataset size.

indicating both the independence between different criteria and the commonality present among high-quality texts.

3.3 Tagging

The quality of the pretraining corpus is initially assessed through rating criteria. However, these criteria alone are insufficient for ensuring diversity in the pretraining samples and for the fine-grained selection of data. Tagging pretraining data into a broad spectrum of topics and fields can ensure diversity within the training corpus. Furthermore, a structured tagging system facilitates the targeted enhancement of the model by incorporating data that address specific areas, consequently improving the model's performance in particular domains. Next, we introduce our hierarchical tagging system.

Tags Design. To systematically categorize the pretraining dataset, we first clearly define 21 primary categories that cover a wide range of human knowledge, from Natural Sciences to Social Events. We then expand this framework by engaging GPT-4, which serves as a human expert, in a two-step iterative dialogue process. The first dialogue iteration yields 255 second-level tags. For the third-level tags, we inform GPT-4 of each first-level category along with its corresponding second-level

tags, prompting the model to generate a total of 793 specific third-level tags under the second-level categories. The details and prompts are in Appendix A.2.

Analysis. We present the result of the tag tree in Figure 10 and the word cloud of the tag tree in Figure 4. To access the tag prediction performance, we manually re-annotated the existing validation split set with tags at the first, second, and third levels. We then compare the accuracy of DecorateLM and GPT-4 using this newly re-annotated validation set. As shown in Table 1, DecorateLM achieves performance comparable to that of GPT-4.

3.4 Editing

The process of rating and tagging extracts valuable data from the pretraining corpus. Despite undergoing a rigorous cleaning pipeline, even high-quality data sourced from the internet may still retain some noise. Inspired by the work of (Maini et al., 2024), we propose to enhance the utilization of this high-quality data by rephrasing it based on the intrinsic attributes of the samples. By transforming the data into different verbal forms, we aim to preserve the core information diversity of the pertaining stage while being as clean as the SFT-stage dataset.
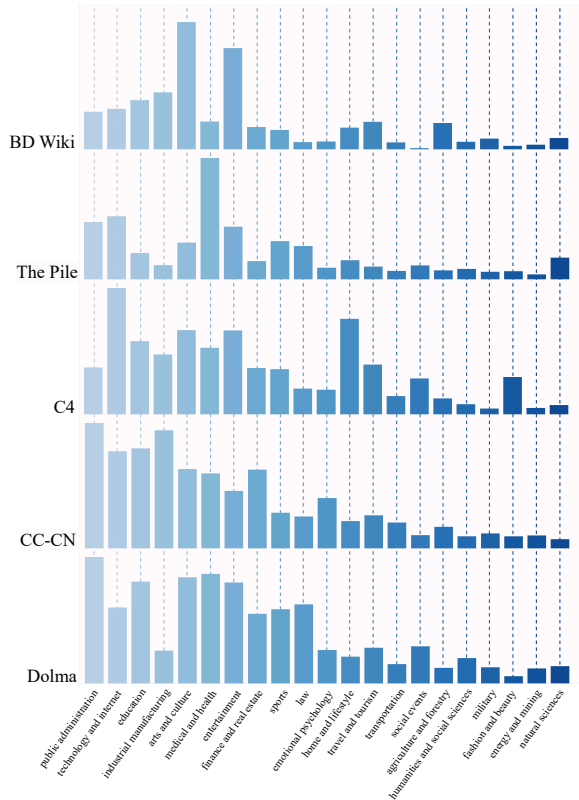
Figure 6: Distribution of first-level tags across different datasets, arranged in descending order by frequency in the decorated corpus.
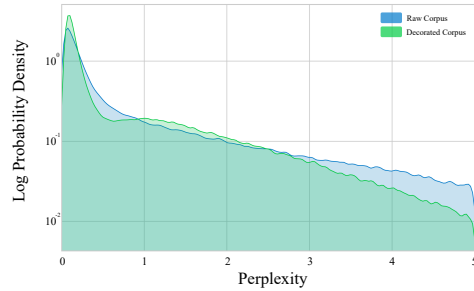

Figure 7: Perplexity distribution of the corpus.
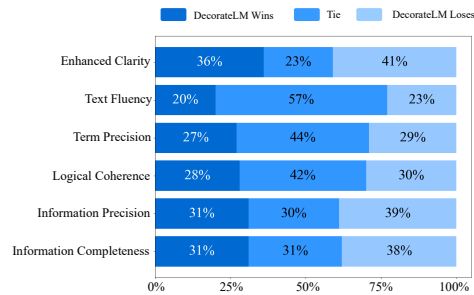
Figure 8: Human Preference for Edited Texts on Validation Set: DecorateLM vs. GPT-4.

Annotated Dataset Construction. We begin by selecting 10,000 data samples, each containing between 50 and 2048 tokens, to create a noisy dataset. We observe that this noisy dataset continues to exhibit issues such as unclear expressions, lack of natural language fluency, and mixed topics that are not fully resolved by standard cleaning methods. This noisy dataset is rephrased using GPT-4 based on prompts in Appendix A.3.

Analysis. Due to the absence of a comprehensive metric for evaluating rephrased text against the original text, we design several custom metrics and use human evaluation to quality-check the rephrased texts. For each evaluation metric, we compare the rephrased outputs of DecorateLM and GPT-4, with human annotators rating each output as a win, lose, or tie. The evaluation metrics are as follows: *Enhanced Clarity*, which determines the text's increased conciseness and clearer expression; *Text Fluency*, which assesses the smoothness and readability of the text; *Term Precision*, which checks the retention of specialized terminology; *Logical Coherence*, which examines the consistency of causal and logical relationships within the

text; *Information Precision*, which verifies that the original meaning, core information, and arguments are accurately preserved; *Information Completeness*, which ensures that no crucial information is missing from the text. The validation set size is 500. As presented in Figure 8, the editing model of DecorateLM, demonstrates satisfactory performance in this task.

3.5 The Final Decorated Corpus

After we train the DecorateLM on the curated annotated dataset, we proceed to decorate the pre-training corpus. Specifically, we select five large pre-training datasets including Common Crawl Chn (CC-CN), Dolma, C4, The Pile, and Baidu Wiki (BD-Wiki). Due to limited resources, we only sample a volume of 100 billion tokens from these datasets.

For the rated and tagged corpus, as shown in Figure 5, the English datasets, Dolma and The Pile, exhibit relatively high ratings and low cross-entropy, making them relatively ideal training corpora that are high-quality and well-balanced across domains. In contrast, the Chinese datasets, BD-wiki and CC-CN, exhibit lower ratings and higher cross-entropy, indicating shortcomings in overall quality and data distribution. This also underscores the necessity of using DecorateLM to improve the quality of the non-English corpus. For the tagging result alone,

the analysis of the distribution of these datasets across the first-level labels is illustrated in Figure 6. Regarding the effectiveness of editing on the Decorated Corpus, the original and edited texts are assessed using the perplexity metric with the CCNet model (Wenzek et al., 2019). The results, shown in Figure 7, indicate a significant reduction in perplexity following the editing process. This improvement suggests that the editing effectively organizes the data in a manner that is more conducive to learning by models, ensuring enhanced comprehensibility and learnability.

4 Experiments

In this section, we conduct data experiments to demonstrate the effectiveness of decorated corpus.

4.1 Experiment Setup

We train the same SLM, MiniCPM-1.2B, used as the backbone for DecorateLM, aiming to improve its performance. MiniCPM-1.2B follows the multi-stage training pipeline (Hu et al., 2024). The stable training stage utilizes a constant learning rate until the decay stage, where the learning rate decreases rapidly. During the decay stage, the loss reduction accelerates significantly. This stage is deemed suitable for ablation studies on different data due to its substantial loss reduction and short training duration. We leverage the last checkpoint before the decay stage to reprocess the decay with both the raw and decorated corpora. Performance is evaluated against a wide range of publicly available benchmarks.

4.2 Experiments on Rating

Given the rating of each test sample, we can select each sample with a probability determined by these ratings (Wettig et al., 2024). We explore two sampling methods.

The first method, referred to as "Separate Criterion Sampling", follows the approach proposed by (Wettig et al., 2024). Specifically, each criterion is given a weight that represents its relative importance. The sampling method begins from the criterion with the highest weight to the lowest one. The transition between criteria happens when the sampled data from the dimension satisfies its predetermined corpus proportion. Within each criterion, data is sampled according to the following weight 1. The ratings for the i-th data point in t-th criterion

are calculated using the following equation:

$$W_{i,t} = e^{\frac{\text{score}_{i,t} - \lambda}{\tau}},\qquad(1)$$

where i is the data point index and t is the criterion index, both λ and τ are set to 50.

The second method, called "Aggregate Criterion Sampling", calculates the sampling weight W_i for the i-th data as follows:

$$W_i = \sum_{t=1}^{8} k_t \cdot e^{\frac{\text{score}_{t,i} - \mu_t}{\sigma_t}},\qquad(2)$$

where the parameter k_t represents the relative significance of each rating dimension.

For both Rat. (Sep.) with weights and Rat. (Agg.) with k_t, the main method assigns a weight of 0.2 to the dimensions of Educational Value, Expertise, Fact and Trivia, and Reasoning Level, while the four remaining dimensions are each assigned a weight of 0.05 according to the authors' prior knowledge of the data quality.

In practice, we sample 58.5B tokens but only use 45B tokens among them as the high-quality data. This has a similar effect as increasing the temperature of sampling in (Wettig et al., 2024).

4.3 Experiments on Tagging

We enhance the diversity and balance of different domains by incorporating a sampling strategy among tags. Intuitively, a large domain should be undersampled and a rare domain should be upsampled. Specifically, we sample an instance with a hierarchical tag of $a \rightarrow b \rightarrow c$ with the weight of

$$W_{a,b,c} = \frac{N_{\text{I}=a}^{\alpha}}{\sum_{i=1}^{N_{\text{I}}} N_{\text{I}=i}^{\alpha}} \cdot \frac{N_{\text{I}=a,\text{II}=b}^{\beta}}{\sum_{i=1}^{N_{\text{I}=a,\text{II}}} N_{\text{I}=a,\text{II}=i}^{\beta}} \cdot \\ \frac{N_{\text{I}=a,\text{II}=b,\text{III}=c}^{\gamma}}{\sum_{i=1}^{N_{\text{I}=a,\text{II}=b,\text{III}}} N_{\text{I}=a,\text{II}=b,\text{III}=i}^{\gamma}},\qquad(3)$$

where $N_{X=x}$ represents the number of instance whose belong to tag x at tag level X. The exponents α, β, γ are similar to what is suggested by (Lample and Conneau, 2019) to tune the distribution to be smooth or concentrated.

For the combined method of Rat. (Agg) & Tag. , we calculate the sampling weights by multiplying the weights of Rat. (Sep.) and Tag..

Domain Coverage Criterion (Avg. (DC)). To demonstrate the improvements brought by making the domain more balanced through tagging, we construct a domain coverage criterion

Method	C-Eval (0-shot)	CMMLU (5-shot)	AGI. (5-shot)	MMLU (5-shot)	Human. (0-shot)	MBPP (0-shot)	GSM. (0-shot)
Base.	47.4	46.8	20.8	45.8	26.2	27.7	38.9
Tag.	47.8 ↑0.4	46.8	21.3 ↑0.5	47.3 ↑1.5	27.4 ↑1.2	28.4 ↑0.7	40.0 ↑1.1
Rat. (Sep.)	45.2 ↓2.2	45.4 ↓1.4	26.4 ↑5.6	46.0 ↑0.2	28.1 ↑1.9	29.1 ↑1.4	41.8 ↑2.9
Rat. (Agg.)	**49.1** ↑1.7	47.0 ↑0.2	26.3 ↑5.5	46.9 ↑1.1	25.6 ↓0.6	30.3 ↑2.6	42.5 ↑3.6
Rat. (Agg.)&Tag.	48.0 ↑0.6	**47.9** ↑1.1	25.3 ↑4.5	46.0 ↑0.2	28.7 ↑2.5	28.1 ↑0.4	40.9 ↑2.0
Edit.	46.7 ↓0.7	47.1 ↑0.3	23.8 ↑3.0	46.9 ↑1.1	27.4 ↑1.2	30.4 ↑2.7	40.1 ↑1.2
Rat. (Agg.)&Edit.	48.1 ↑0.7	47.8 ↑1.0	**28.0** ↑7.2	47.5 ↑1.7	**31.7** ↑5.5	30.0 ↑2.3	**42.6** ↑3.7
Rat. (Agg.)&Tag.&Edit.	47.4	46.4 ↓0.4	24.3 ↑3.5	**47.6** ↑1.8	29.3 ↑3.1	**30.9** ↑3.2	40.3 ↑1.4

Method	MATH (4-shot)	BBH (0-shot)	ARC-E (0-shot)	ARC-C (0-shot)	Trivia. (0-shot)	Avg. (DC)	Avg.
Base.	3.5	28.5	78.2	61.8	6.0	37.5	36.1
Tag.	4.6 ↑1.1	27.8 ↓0.7	79.2 ↑1.0	62.1 ↑0.3	12.7 ↑6.7	41.8 ↑4.3	37.5 ↑1.4
Rat. (Sep.)	6.5 ↑3.0	28.4 ↓0.1	78.8 ↑0.6	61.4 ↓0.4	10.4 ↑4.4	39.2 ↑1.7	37.4 ↑1.3
Rat. (Agg.)	4.8 ↑1.3	28.5	79.3 ↑1.1	**63.0** ↑1.2	15.6 ↑9.6	41.1 ↑3.6	38.5 ↑2.4
Rat. (Agg.)&Tag.	**6.7** ↑3.2	28.0 ↓0.5	78.8 ↑0.6	62.6 ↑0.8	13.7 ↑7.7	43.1 ↑5.6	38.3 ↑2.2
Edit.	5.6 ↑2.1	29.2 ↑0.7	77.8 ↓0.4	62.0 ↑0.2	22.0 ↑16.0	40.5 ↑3.0	38.4 ↑2.3
Rat. (Agg.)&Edit.	4.3 ↑0.8	32.7 ↑4.2	**79.5** ↑1.3	62.7 ↑0.9	24.9 ↑18.9	42.8 ↑5.3	**40.2** ↑4.1
Rat. (Agg.)&Tag.&Edit.	5.5 ↑2.0	29.8 ↑1.3	77.9 ↓0.3	**63.0** ↑1.2	27.8 ↑21.8	**45.0** ↑7.5	39.6 ↑3.5

Table 2: Comparison of benchmark performance across different strategies.

by averaging the accuracy scores of 6 tasks within the following 5 domains. *Sports* domain is represented by SportQA (Xia et al., 2024) dataset. *Medicine* domain is represented by MedM-CQA (Pal et al., 2022) and MedQA-USMLE (Jin et al., 2021) datasets. *Law* domain is represented by JECQA (Zhong et al., 2020) dataset. *Natural sciences* domain is represented by SciQ (Welbl et al., 2017) dataset. *Finance* domain is represented by OpenFinData dataset[1].

4.4 Experiments on Editing

Building upon the existing methods (Baseline, Rat. (Agg.), and Rat. (Agg.)&Tag.), we introduce the Editing approach. We randomly select one-third of the training data to be replaced with edited data.

4.5 Results

In this section, we present the results of data experiments. Details and specific settings of the evaluation experiments can be found in Appendix D.

As shown in Table 2, the integration of various methods yields several significant insights:

- **Rating:** Both rating sampling methods exhibit superior overall performance compared

to the baseline. Rat. (Agg.) improves almost all tasks and achieves an overall average score increase of 2.4 points, which is greater than Rat. (Sep.).

- **Tagging:** The Tag. method shows a slight improvement over the baseline in overall benchmarks and achieves a significant 4.3-point increase on the Domain Coverage benchmark. The Rat. (Agg) & Tag. method has comparable overall performance to Rat. (Agg), with an additional 2-point improvement on Avg.(DC). Moreover, to validate the effectiveness of domain filtering, we evaluate an MMLU-oriented tagging model, as depicted in Figure 9. The model targets 20 specific MMLU subtasks, enhancing their sampling probability. It demonstrates improvement in 15 of these 20 tasks compared to the Tag. method, thereby affirming the efficacy of the tagging system in modifying domain composition for targeted reinforcement.

- **Editing:** Integration of the Editing method significantly enhances model performance on downstream tasks. Edit. increases the average score by 2.3 percentage points compared to the baseline, demonstrating its effectiveness

in rephrasing training data.

- **Rating and Editing:** Rat. (Agg.)&Edit. emerges as the best-performing method, enhancing the average score by 4.1 points relative to the baseline and demonstrating improvements across all tasks. Rat. (Agg.)&Tag.&Edit. attains the highest score on Avg. (DC) and maintains excellent performance in other tasks, suggesting that the integration of tagging with rating and editing expands the models' knowledge base without substantially compromising depth.

5 Conclusion

In this paper, we present *DecorateLM*, a data engineering method designed to refine the pre-training corpus through data rating, tagging and editing. *DecorateLM* employs a dual-training strategy, wherein two student models with 1.2 B parameters are trained: one designed for rating and tagging, and the other focused on editing. Our experiments show that introducing rating and editing in data corpus significantly enhances data quality by improving the overall performance of SLM on various existing benchmarks. Furthermore, our empirical study verifies that the implemented tagging strategy achieves a more balanced distribution of categories within the training dataset. This equilibrium in categorization enables a more thorough comprehension of SLM proficiency across diverse domains. These encouraging results underscore the importance of training data quality in fully exploiting the capabilities of Large Language Models, thereby suggesting several compelling avenues for future research.

6 Limitations

Our study, while enhancing the quality of data effectively, is subject to several limitations. Firstly, the biases present in GPT-4 may be reflected in the fine-tuning data used for DecorateLM, potentially causing DecorateLM to inherit these biases Additionally, due to computational and time constraints, we limit our model training to 1.2 billion parameter models using high-quality data. The generalizability of our findings would benefit from replication with larger language models and a wider range of datasets. Thirdly, our investigation is confined to training models during the decay stage using the Decorated Corpus. An additional dimension to our work would involve creating a dataset of 1.1 trillion tokens with DecorateLM, followed by training a model from scratch on this enlarged dataset, which we believe represents an important direction for future research.

Moreover, although DecorateLM performs well in filtering data from large-scale web data, its ability to handle more specialized domains still requires improvement. The classification and labeling of the diverse content of the real world by humans are challenging to fully capture with a three-layer labeling system. Future research could explore a more granular labeling system to enhance the model's precision and breadth in professional fields. Lastly, while DecorateLM considered both English and Chinese, it did not take other languages such as French and Russian into account, which may limit its generalizability to other languages.

An additional limitation lies in the current approach to sampling, which may not adequately capture the nuanced relationships between ratings and taggings across various tasks. Therefore, future research should explore a wider array of sampling strategies for rating and tagging to assess their impact on task performance more comprehensively.

7 Ethical Considerations

As we develop DecorateLM, we recognize the inherent risk of introducing or magnifying biases within our datasets. The training process, while intended to refine and improve data accuracy, could inadvertently perpetuate biases present in the original data. This raises significant ethical concerns, as biased data can lead to unfair outcomes in decision-making processes that rely on our enhanced training data.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, et al. 2023. Alpagasus: Training a better alpaca with fewer data. *arXiv preprint arXiv:2307.08701*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. 2023. Mods: Model-oriented data selection for instruction tuning. *arXiv preprint arXiv:2311.15653*.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

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

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

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. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Teven Le Scao, Angela Fan, Christopher Akiki, El- lie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b- parameter open-access multilingual language model.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Bald- win. 2023a. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023b. Textbooks are all you need ii: phi-1.5 techni- cal report. *arXiv preprint arXiv:2309.05463*.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. 2023. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. *arXiv preprint arXiv:2312.15685*.

Shayne Longpre, Gregory Yauney, Emily Reif, Kather- ine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2023. A pretrainer's guide to training data: Measur- ing the effects of data age, domain coverage, quality, & toxicity. *arXiv preprint arXiv:2305.13169*.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Jun- yang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2023. # instag: Instruction tagging for analyz- ing supervised fine-tuning of large language models. In *The Twelfth International Conference on Learning Representations*.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient lan- guage modeling. *arXiv preprint arXiv:2401.16380*.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I Jor- dan, et al. 2018. Ray: A distributed framework for emerging {AI} applications. In *13th USENIX sym- posium on operating systems design and implementa- tion (OSDI 18)*, pages 561–577.

Ankit Pal, Logesh Kumar Umapathi, and Malaikan- nan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical do- main question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei,

and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software de- velopment. *arXiv preprint arXiv:2307.07924*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the lim- its of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bo- gin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three tril- lion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits rea- soning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, Chenxia Li, Liu Yang, Xilin Luo, Xuejie Wu, Lunan Liu, Wenjun Cheng, Peng Cheng, Jianhao Zhang, Xiaoyu Zhang, Lei Lin, Xi- aokun Wang, Yutuan Ma, Chuanhai Dong, Yanqi Sun, Yifu Chen, Yongyi Peng, Xiaojuan Liang, Shuicheng Yan, Han Fang, and Yahui Zhou. 2023. Skywork: A more open bilingual foundation model. *Preprint*, arXiv:2310.19341.

Johannes Welbl, Nelson F Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. *arXiv preprint arXiv:1707.06209*.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Con- neau, Vishrav Chaudhary, Francisco Guzmán, Ar- mand Joulin, and Edouard Grave. 2019. Ccnet: Ex- tracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, et al. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arXiv preprint arXiv:2110.04725*.

Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuan-fang Wang, and Weining Shen. 2024. Sportqa: A benchmark for sports understanding in large language models. *arXiv preprint arXiv:2402.15862*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems*, 36:34201–34227.

Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. Cluecorpus2020: A large-scale chinese corpus for pre-training language model. *ArXiv*, abs/2003.01355.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Jecqa: a legal-domain question answering dataset. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9701–9708.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

A Full Prompts

A.1 Prompts of Rating

Prompt Template

Compare which text {criterion}
Your judgement should not be influenced by the language the text is written in, the length of the text and the order in which the texts are presented.
If the texts have similar quality, you should still make a relative judgement and choose the label of the preferred text.
You must respond with format:
"Choice: 1 or 2\nWhy: reason of choice"

Text 1: ... {text_1} ...

Text 2: ... {text_2} ...

Now you have to choose between either 1 or 2. Note that respond only with the format mentioned.

Educational Value

has more educational value. It has more educational value if it includes clear explanations, step-by-step reasoning, or detailed concepts which is clear enough for children to understand.
Prefer text which has more detailed ideas or explanations which is sufficiently clear to convey them to a child.

Expertise

requires greater expertise and deeper prerequisite knowledge to understand it.
For example, "The relativistic Dirac equation, which combines principles of quantum mechanics and special relativity, predicts the existence of antimatter and elucidates the intrinsic spin of fundamental particles." requires great physics expertise to understand.

Fact and Trivia

contains more facts and trivia. The facts and trivia should be accurate.
Prefer text which have more number of facts. Put lower priority to facts which contain mathematical calculations and with too deep "concepts and explanations" .

Reasoning Level

has higher reasoning level. It has high reasoning level when it requires more reasoning, logical and mathematical thinking skills or chain of thought thinking.

Scarcity

is more relatively unknown. It should be truthful and little known to the general public.
Prefer unpopular accurate facts over fictional stories.

Structural Format

has better structural format. It has better structural format when it has a well-defined structure such as outline format, Markdown, numbered list, bulleted list, JSON, table format, headings and subheadings format or other organizational templates.
First, consider the visual structure of text. Then, only consider the content or logical flow of text.

Story-likeness

is more likely to be a story. It is more like a story when it narrates a story or it describes a scene or situation in details.

Subjectivity

contains more subjectivity, e.g, it includes more subjective perspectives, opinions, personal views or feelings. Avoid choosing text which conveys objective, factual and widely accepted, accurate knowledge. Prefer text which personal opinions such as dialogues or feelings over text which seems like a formal examination question and answer.

Generate Structural Format Data

You are tasked with generating text data that has clear and organized formatting structures. Some structural format are list, markdown, headings and subheadings, table, json, html, xml, latex, columnar formats etc.
The data should maintain a coherent structure with organized sections, numbering, tables, code formatting, hierarchical structure, outlines or other organizational templates where appropriate. You should not include all of the formats in one data. One data can mix of one, two or three formats.
You can add various knowledge and facts into data to make data more informative and longer.
Please generate 3 lengthy and informative examples about '<topic>' showcasing different formatting styles and content. Split examples with <split>

A.2   Prompts of Tagging

Prompt Template For Summary

Your objective is to summarize the provided text: [begin] {instance} [end], within 100 words, including the relevant information for the use case in the summary as much as possible.
The summary will represent the input data for clustering in the next step.
Be concise and clear.
Do not add phrases like "This is the summary of" or "Summarized text:"...
Do not include any line breaks in the summary.
Provide your answer in English only.
Your comprehensive output should mirror this structure: {{"summary": ""}}.

Prompt Template For First-level Tagging

You are an advanced tagging system designed to identify the most pertinent theme within a given text passage: [begin] {instance} [end].
Your role is to analyze the text meticulously and choose the most fitting tag from the predefined list: Natural Sciences, Humanities and Social Sciences, Industrial Manufacturing, Medical and Health, Agriculture and Forestry, Energy and Mining, Finance and Real Estate, Education, Transportation, Technology and Internet, Law, Military, Travel and Tourism, Entertainment, Arts and Culture, Emotional Psychology, Fashion and Beauty, Sports, Home and Lifestyle, Public Administration, and Social Events.
Your task is to determine the single most relevant tag that encapsulates the primary theme of the text.
Your selection should be substantiated with a detailed explanation, elucidating why this tag is the most accurate representation of the text's central subject matter.
Your output should follow this structure: {{"tag": "Selected Tag", "explanation": "Provide a detailed explanation in English on why this is the most fitting tag."}}.

Prompt Template For Second-level And Third-level Tagging

You are an advanced tagging system designed to categorize a given text passage related to the first level tag "{first_level_tag}" into specific second and third-level tags within a predefined hierarchy.
Here is the tag hierarchy for the "{first_level_tag}" category in json format: {tag_tree}
Here is the given text passage: [begin] {instance} [end].
Your task is to analyze the text snippet above and assign the most fitting second-level and third-level tags, ensuring both tags align within the same hierarchical path.
The output should precisely reflect the main focus of the text, justifying why these tags are the most suitable choices.
Your output should follow this structure: {{"second_level_tag": "Selected Second Level Tag", "third_level_tag": "Selected Third Level Tag", "explanation": "Provide a detailed explanation in English on why these tags accurately represent the text's core content."}}.

A.3   Prompts of Editing

Editing Template

For the following paragraph give me a diverse paraphrase of the same in high quality language as in sentences on Wikipedia. Generate text directly from the provided content. Do not exceed the original information or add explanations.
text:

B DecorateLM Training

B.1 Details of rating and tagging model

We employ MiniCPM-1.2B (Hu et al., 2024) as our base model. Utilizing the previously proposed rating and tagging methodologies, we collect rating and three-level tagging of 30,000 training data samples and subsequently apply supervised fine-tuning to the MiniCPM-1.2B with a learning rate of 0.00125 and total batch size of 480 every iteration. The fine-tuning process is conducted on three machines, each equipped with eight Nvidia A100 GPUs. We implement an decay step every 120 iterations and a warm-up phase of 3 iterations, yielding distilled rating and tagging models. We observe that only 200 steps are needed to fine-tune the model to its optimal performance in rating and tagging.

B.2 Details of editing model

Similar to the rating and tagging model, we utilize the previously proposed editing method and collect 10,000 data samples with rephrased content by GPT-4. Subsequently, we apply supervised fine-tuning to MiniCPM-1.2B with the same method and hyperparameters as the rating and tagging model, yielding an editing model. We observe that fine-tuning the model for optimal performance in editing tasks requires 600 steps, a notably higher number compared to the steps needed for the rating and tagging model. This increased demand for training iterations likely reflects the greater complexity and difficulty associated with editing tasks.

C Further Analysis of DecorateLM

C.1 Cost Analysis

Utilizing the vLLM framework (Kwon et al., 2023) and Ray (Moritz et al., 2018), we facilitate the generation of synthetic data across distinct phases with varying processing efficiencies on a single Nvidia A100 GPU. In the rating and tagging phase, the MiniCPM-1.2B model processes 16 million tokens per hour, requiring approximately 6,250 GPU hours to generate 100 billion tokens. Conversely, in the editing phase, the same model configuration processes 12.5 million tokens per hour, necessitating around 8,000 GPU hours for the production of an equivalent volume of tokens.

C.2 Details of Decorated Corpus

The Decorated Corpus is constructed from a variety of datasets, each contributing to the total com-

position according to the proportions specified in Table 3.

Dolma. Dolma dataset (Soldaini et al., 2024) encompasses a comprehensive corpus designed for advancing the field of language model pretraining.

CC-CN. CC-CN dataset is composed of a combination of sources from (Xu et al., 2020), (Wei et al., 2023), and (Wu et al., 2021)

C4. C4 dataset (Raffel et al., 2020) represents a significant milestone in the field of natural language processing, particularly within the domain of transfer learning.

The Pile. The Pile dataset (Gao et al., 2020) is a substantial contribution to large-scale language model training, featuring an extensive corpus of 825 GiB of English text.

BD Wiki. The BD Wiki dataset, derived from the Baidu Baike[2], is a semi-open Chinese online encyclopedia operated by Baidu Inc.

D Training With Decorated Corpus

D.1 Experimental Details

We employ the pre-decay version of MiniCPM-1.2B, pre-trained on a corpus comprising 800 billion tokens, as our base model. For training, the Decorated Corpus and additional high-quality datasets are utilized. The base model undergoes a decay process over 20,000 steps with a learning rate of 0.01 and a batch size of 1200 tokens per iteration, distributed across 10 machines, each equipped with eight A100-80GB GPUs. A decay step is implemented every 5000 iterations.

D.2 Evaluation Details

The overall evaluation utilizes the open-source tool UltraEval[3]. The underlying inference and acceleration use the open-source framework vLLM (Kwon et al., 2023), and the dataset includes commonly used datasets: C-Eval (Huang et al., 2024) and CMMLU (Li et al., 2023a) for Chinese knowledge, AGI-Eval (Zhong et al., 2023) for World Knowledge, MMLU (Hendrycks et al., 2020) for English knowledge, HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) for coding, GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) for mathematics,

---

[2] https://baike.baidu.com/
[3] https://ultraeval.openbmb.cn/home

Dataset	Dolma	CC-CN	C4	The Pile	BD Wiki
# Tokens (millions)	320	290	200	100	90

Table 3: Composition of the Decorated Corpus Dataset.

Method	Sport. (0-shot)	MedMC. (0-shot)	Med.-US. (0-shot)	JEC. (0-shot)	SciQ (0-shot)	OpenFin. (0-shot)	Avg. (DC)
Base.	16.5	29.8	28.0	31.4	71.3	48.1	37.5
Tag.	20.9 ↑4.4	36.9 ↑7.1	34.4 ↑6.4	35.4 ↑4.0	74.0 ↑2.7	**48.9** ↑0.8	41.8 ↑4.3
Rat. (Sep.)	7.0 ↓9.5	36.8 ↑7.0	36.6 ↑8.6	35.4 ↑4.0	77.2 ↑5.9	42.3 ↓5.8	39.2 ↑1.7
Rat. (Agg.)	15.0 ↓1.5	36.9 ↑7.1	37.1 ↑9.1	34.5 ↑3.1	77.4 ↑6.1	45.7 ↓2.4	41.1 ↑3.6
Rat. (Agg.)&Tag.	22.2 ↑5.7	**39.9** ↑10.1	36.3 ↑8.3	36.4 ↑5.0	78.4 ↑7.1	45.2 ↓2.9	43.1 ↑5.6
Edit.	16.8 ↑0.3	33.0 ↑3.2	32.1 ↑4.1	**36.6** ↑5.2	75.9 ↑4.26	48.7 ↑0.6	40.5 ↑3.0
Rat. (Agg.)&Edit.	17.5 ↑1.0	36.9 ↑7.1	39.5 ↑11.5	36.5 ↑5.1	80.5 ↑9.2	45.6 ↓2.5	42.8 ↑5.3
Rat. (Agg.)&Tag.&Edit.	**25.8** ↑9.3	38.8 ↑9.0	**40.1** ↑12.1	36.4 ↑5.0	**80.7** ↑9.4	48.1	**45.0** ↑7.5

Table 4: Comparison of rare domain benchmark performance across different strategies.

and BBH (Srivastava et al., 2022) for logic reasoning, and ARC-E (Clark et al., 2018), ARC-C (Clark et al., 2018)for commonsense reasoning, and TriviaQA (Joshi et al., 2017) for Reading Comprehension. Additionally, we conduct the Domain Coverage (DC) benchmark to evaluate the model's capability across various domain-specific knowledge bases. The DC Benchmark includes datasets such as SportQA (Xia et al., 2024) for sports, MedMCQA (Pal et al., 2022) and MedQA-USMLE (Jin et al., 2021) for medicine, JECQA (Zhong et al., 2020) for law, SciQ (Welbl et al., 2017) for natural sciences, and OpenFinData[4] for finance.

E   Inspecting cases of DecorateLM

Case 1

Text:
Photodynamic diagnosis (PDD) is available at some centres. PDD makes cancer cells visible under violet light. This makes abnormal growths (tumours) easier to find and remove. More complete removal reduces the risk of cancer coming back.PDD is used in urology to test for bladder cancer. A catheter is inserted through the urethra and into the bladder. A solution of 5-aminolaevulinic acid or hexaminolaevulinic acid is passed through the catheter into the bladder. Then the catheter is taken out immediately. Cancer cells in the bladder process the chemicals in the solution and become fluorescent under violet light. No side-effects or complications have been reported for PDD.
Rating:

Educational Value:	99	Expertise:	92
Fact and Trivia:	92	Scarcity:	85
Reasoning Level:	93	Story-likenss:	55
Structural format:	91	Subjectivity:	25

Tagging:
First Level Tag: Medical And Health
Second Level Tag: Medical Procedures
Third Level Tag: Diagnostic Procedures
Editing:
Photodynamic diagnosis (PDD) is accessible at certain facilities, a technique that enables the visualization of cancer cells under violet light, thereby enhancing the detection and removal of irregularities, such as tumours, with increased precision. This method further minimizes the likelihood of cancer cells recurring. In the field of urology, PDD is utilized for the identification of bladder cancer. The procedure involves the insertion of a catheter through the urethra into the bladder. Within the catheter, a solution containing 5-aminolaevulinic acid or hexaminolaevulinic acid is delivered. Following this, the catheter is promptly removed. It is observed that cancer cells within the bladder respond to these chemicals, converting them into fluorescent particles under the influence of violet light. Due to its safe and effective operation, no adverse effects or issues have been noted associated with PDD.

Case 2

Text:
Tron (TRX) is an ambitious project which aims to own the bragging rights regarding the entertainment industry. To make that a reality, a day hardly ends...Tron (TRX) is undoubtedly an ambitious project in the crypto community, and it aims to own the bragging rights of the entertainment industry. However, they are...Before I start explaining how Tron (TRX) Mainnet, Odyssey, is enhancing the chances of Tron adoption already, I need to explain some crucial thoughts with you....What will be the

[4]https://github.com/open-compass/OpenFinData

price of Tron (TRX) in five years from now? Tron is one of the top cryptocurrencies in the world (always in the spotlight). The cryptocurrency's mainnet is now live, and on its independence day, it...Pornhub, one of the largest X-rated video sites, expands its virtual currency payment options announcing they will be soon accepting Tron (TRX). In recent times, Tron...Can Tron (TRX) become one of the top 5 cryptos in the market now?In recent days, Tron (TRX) has been experiencing one of the biggest tractions of the market, a fact that without a doubt is helping the token...

Rating:

Educational Value:	45	Expertise:	54
Fact and Trivia:	80	Scarcity:	10
Reasoning Level:	51	Story-likenss:	60
Structural format:	55	Subjectivity:	80

Tagging:

First Level Tag:	Technology and Internet
Second Level Tag:	Entertainment and Media
Third Level Tag:	Digital Media

Editing:

Tron (TRX) embodies a visionary endeavor to dominate the boundaries of the entertainment sector, a goal that promises to be achieved on a nearly perpetual basis. In pursuit of this ambitious objective, Tron (TRX) continually strives to enhance its presence. In the realm of cryptocurrency, Tron (TRX) stands out for its bold ambitions, vying for control over the entertainment industry's prestige. Despite its aspirations, Tron (TRX) is already making strides in securing a larger user base, thanks to its innovative Mainnet, Odyssey. This initiative is pivotal in laying the groundwork for Tron's broader adoption. As we reflect on the potential future value of Tron (TRX) in the coming five years, it is noteworthy that Tron holds a prominent position among the world's cryptocurrencies, drawing significant attention. The cryptocurrency's flagship network has commenced operations, and its founding day marks an important milestone for its development. The entertainment sector's leading platform, Pornhub, announces a expansion into virtual currency payments, revealing plans to begin accepting Tron (TRX) as a payment method shortly. This move highlights the growing acceptance and adoption of Tron (TRX) within the digital landscape. Furthermore, there exists a discussion surrounding the potential for Tron (TRX) to surpass other top 5 cryptos in the market, with recent market trends clearly indicating its robust growth and widespread acceptance.

Kimmy belt, $625, at Isabel Marant, Design District; printed velvet trousers, $900, by Paco Rabanne at The Webster, Bal Harbour Shops and South Beach; Ellabrita strass sandal 105, $1,150, by René Caovilla at Neiman Marcus, Bal Harbour Shops and Shops at Merrick Park; diamond earrings, price upon request, at elanjewels.us. Gown, $25,000, Valentino, Bal Harbour Shops and Design District; feather boa, $3,990, Loewe, Design District.

Rating:

Educational Value:	10	Expertise:	2
Fact and Trivia:	51	Scarcity:	1
Reasoning Level:	11	Story-likenss:	50
Structural format:	36	Subjectivity:	63

Tagging:

First Level Tag:	Fashion and Beauty
Second Level Tag:	Fashion Apparel
Third Level Tag:	Formal Attire

Editing:

A gown from Valentino, Bal Harbour Shops and Design District, retails for $12,900. Additionally, diamond stud earrings are available at a custom price, with the source being elanjewels.us. Jennifer Hudson is celebrated for her profound emotional depth, capturing the essence of her characters and elevating musical pieces to new heights. Her innate talent, which has attracted a multitude of fans since she was a young age, is poignantly depicted in her latest cinematic venture. Erdem's Larkin coat, priced at $6,770, is showcased at Saks Fifth Avenue, alongside other collections in Bal Harbour Shops, Brickell City Centre, and Dadeland Mall. The Bullet bodysuit, priced at $350, features a satin material by Fleur du Mal at Intermix, along with other merchandise in these same locations. A belt by Kimmy, priced at $625, is available at Isabel Marant in the Design District. For a more contemporary look, a printed velvet trouser, priced at $900, by Paco Rabanne is offered at The Webster in Bal Harbour Shops and South Beach. Elenabrita's Ellabrita strass sandal 105, priced at $1,150, is designed by René Caovilla and available at Neiman Marcus, Shops at Merrick Park, and additional retailers. Diamond earrings, once requested, can be purchased from elanjewels.us. A gown from Valentino, priced at $25,000, is available from Bal Harbour Shops and Design District, while a feather boa, priced at $3,990, adds a distinctive touch to Loewe's designs in the Design District.

Case 3

Text:

Gown, $12,900, Valentino, Bal Harbour Shops and Design District; diamond stud earrings, price upon request, elanjewels.us. Jennifer Hudson, emotion is everything. It's how she breathes life into a character. It's how she makes a song explode. And it's why—since Hudson was a child—people are drawn to her talent like a moth to a flame. Well, wait until you see her newest film. Larkin coat, $6,770, by Erdem at Saks Fifth Avenue, Bal Harbour Shops, Brickell City Centre and Dadeland Mall; satin Bullet bodysuit, $350, by Fleur du Mal at Intermix, Bal Harbour Shops, Brickell City Centre and Lincoln Road;
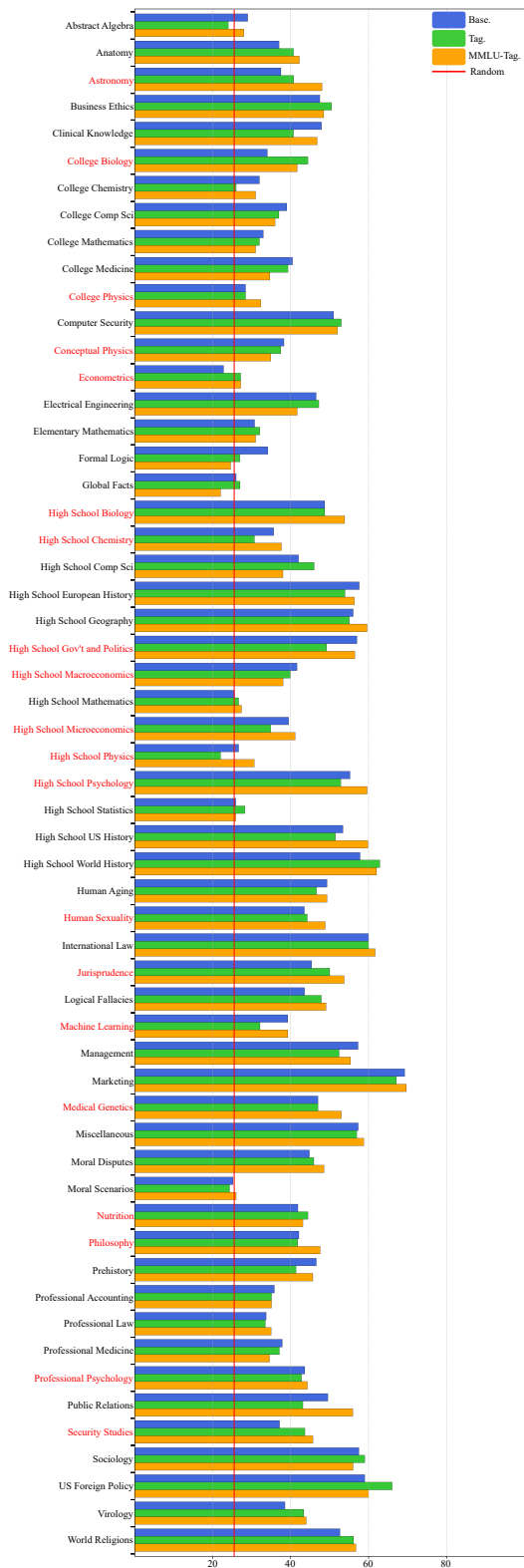
Figure 9: The performance of the MMLU-Tag. Model across the various subtasks of MMLU. The tasks where the sampling weights are increased on the corresponding tags based on the Tag. Methods are highlighted in red.

Figure 10: The tagging tree hierarchy. Only first and second-level tags are shown.