LongWanjuan: Towards Systematic Measurement for Long Text Quality

Xiaoran Liu[1,2*], Kai Lv[1,2*], Qipeng Guo[2], Hang Yan[2], Conghui He[2], Xipeng Qiu[1], Dahua Lin[2]
[1]School of Computer Science, Fudan University, [2]Shanghai AI Laboratory
{klv21,xrliu24}@m.fudan.edu.cn,
{guoqipeng,yanhang,heconghui,lindahua}@pjlab.org.cn
xpqiu@fudan.edu.cn

Abstract

The quality of training data are crucial for enhancing the long-text capabilities of foundation models. Despite existing efforts to refine data quality through heuristic rules and evaluations based on data diversity and difficulty, there's a lack of systematic approaches specifically tailored for assessing long texts. Addressing this gap, our work systematically measures the quality of long texts by evaluating three fundamental linguistic dimensions: coherence, cohesion, and complexity. Drawing inspiration from the aforementioned three dimensions, we introduce a suite of metrics designed to evaluate the quality of long texts, encompassing both statistical and pre-trained language model-based ones. Leveraging these metrics, we present LongWanjuan, a bilingual dataset specifically tailored to enhance the training of language models for long-text tasks with over 160B tokens. In LongWanjuan, we categorize long texts into holistic, aggregated, and chaotic types, enabling a detailed analysis of long-text quality. Furthermore, we devise a data mixture recipe that strategically balances different types of long texts within LongWanjuan, leading to significant improvements in model performance on long-text tasks.

1 Introduction

Effectively processing long texts is a crucial capability of language models and has recently become a focal point of research (Chen et al., 2023; Peng et al., 2023; Liu et al., 2023b). Tasks such as long document summarization (Zhong et al., 2021), long document question answering (Dasigi et al., 2021), repository-level code tasks (Liu et al., 2023a), and retrieval-augmentation generation (Xu et al., 2023) often involve handling thousands or even tens of thousands of tokens.

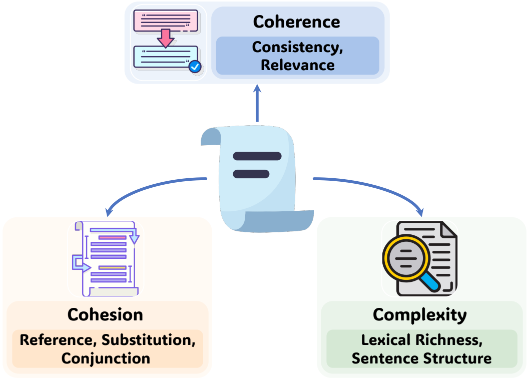

Coherence
Consistency, Relevance
Cohesion
Reference, Substitution, Conjunction
Complexity
Lexical Richness, Sentence Structure

Figure 1: The three dimensions for measuring the quality of long texts: coherence, cohesion and complexity.

The quality of data is vital for the long-text capabilities of foundation models (Zha et al., 2023; Xiong et al., 2023; Rozière et al., 2023). There have been efforts made to improve data quality. Some approaches employ heuristic rules, such as deduplication and the removal of overly short data entries (Soboleva et al., 2023; Penedo et al., 2023). Additionally, some other approaches consider data diversity and perplexity based on pre-trained language models (Tirumala et al., 2023; Marion et al., 2023). However, these filtering rules are designed for general training data and do not take into account the unique characteristics of long texts.

To systematically assess the quality of long texts, we adhere to linguistic fundamentals and evaluate them through three dimensions: coherence (Wang and Guo, 2014), cohesion (Halliday and Hasan, 2014; Carrell, 1982), and complexity (Pallotti, 2015), as illustrated in Figure 1. These three dimensions have often been used to evaluate and analyze text quality(Mathias and Bhattacharyya, 2018). **Coherence** measures the overall consistency and clarity of the text as a whole(Zhong et al.; Wu et al., 2023; Shrivastava et al., 2018; Cho et al.). **Cohesion** gauges the strength of connections between sentences or sections of the text(Zhong et al.; Wu et al., 2023; Cho et al.; Tan et al., 2022). **Complexity** assesses the linguistic sophistication within the

[*] Equal contribution.

Findings of the Association for Computational Linguistics: EACL 2024, pages 5709–5725
November 12-16, 2024 ©2024 Association for Computational Linguistics

text(Imperial and Madabushi, 2023; Li et al., 2022; Salman et al., 2023). Given that long texts typically contain more extensive content, they necessitate elevated levels of these characteristics to effectively convey information and engage in discussion. Drawing from these three fundamental dimensions, we propose a set of metrics to quantitatively analyze the quality of long texts. These metrics encompass both statistical and pre-trained model-based approaches, offering strong interpretability. Further details on these metrics can be found in Section 3.

Based on the characteristics across these three dimensions, we categorize the long texts in pre-training dataset into three types: **holistic long texts**, encompassing complete works such as books, academic papers, reports, novels, and interviews; **aggregated long texts**, consisting of short texts related by topic or fragmented texts like extensive lists or tables; and **chaotic long texts**, characterized by nonsensical content such as garbled data. Drawing upon these classifications, we manually annotated a validation set of 200 samples from SlimPajama (Soboleva et al., 2023) and Wanjuan (He et al., 2023) to validate the correlation between our proposed metrics and human judgments. Our quantitative metrics effectively differentiate between the three categories of long texts.

Building on these analysis and metrics, we create a bilingual long-text dataset with category labels, named LongWanjuan, containing over 160B tokens. With LongWanjuan, we propose a data mixture recipe to mitigate the imbalance between holistic long texts and aggregated long texts within the dataset. Specifically, by removing chaotic long texts and upsampling aggregated long texts, we continue to train InternLM2-7B (Team, 2023) with an additional 5B tokens, thereby achieving state-of-the-art performance for long texts on models of the 7B parameter scale. The effectiveness and generalizability of this recipe are analyzed in Section 5.4.

In summary, our contributions are as follows:

1. To the best of our knowledge, this is the first work to systematically analyze and introduce quantitative metrics for assessing the quality of long texts. Grounded in linguistic principles, we measure the quality of long texts in terms of coherence, cohesion, and complexity.

2. Leveraging SlimPajama and Wanjuan, we constructed a bilingual long-text dataset with over 160B tokens, LongWanjuan, which is available to the community as an open-source resource.

3. Based on LongWanjuan, we devise a data mixture recipe to mitigate the imbalance in the dataset, and advance to a new state-of-the-art long-text model at the 7B parameter scale, demonstrating a 13.07% improvement over the untrained baseline on Longbench (Bai et al., 2023b).

2 Related Work

2.1 Pre-training Data Pruning

The quality of pre-training data plays a crucial role in the performance of foundation models (Rae et al., 2021; Du et al., 2022; Xiong et al., 2023; Rozière et al., 2023; Gunasekar et al., 2023). Several studies have enhanced data quality by pruning the original training data into a subset.

Some works primarily focus on heuristic rules and deduplication to improve data quality. Raffel et al. (2020) and Soboleva et al. (2023) employ similar heuristic rules to enhance data quality, including the removal of overly short entries and deduplication. Abbas et al. (2023) leverages embeddings from pre-trained models to further eliminate semantic duplicates. Another notable contribution is RefinedWeb (Penedo et al., 2023), which meticulously designs a comprehensive data processing pipeline.

Moreover, several studies take into consideration the data diversity and difficulty to prune data. Tirumala et al. (2023) employs clustering-based methods to augment data diversity. Marion et al. (2023) evaluates the effectiveness of perplexity, EL2N (Paul et al., 2021), and memorization score (Biderman et al., 2023) in assessing data difficulty. Maharana et al. (2023) regards data diversity and difficulty as complementary aspects, selecting data through forward and reverse message passing on a dataset graph.

Distinct from these studies that concentrate on general pre-training data, our research specifically targets long texts. It is essential to highlight that our work extends beyond mere data curation and is applicable in a wider range of contexts for evaluating the quality of long texts.

2.2 Text Quality Assessment

Several works score texts through supervised learning. Alikaniotis et al. (2016) trains score-specific

	Low Level Example	**High Level Example**
Coherence	The project aims to reduce carbon emissions by 25% within the next five years. Strawberries are rich in vitamins and antioxidants. It's raining today.	The project **aims to** **reduce carbon emissions** by 25% within the next five years. **This goal** will be achieved through the implementation of **renewable energy sources** and improved **energy efficiency**. **The initiative** reflects our commitment to **environmental sustainability**.
Cohesion	I prepared the soil in my garden. I planted some tomato seeds. I watered seeds in my garden.	**Firstly**, I prepared the soil in my garden. **Then**, I planted some tomato seeds in **the prepared ground**. **After that**, I watered **them**.
Complexity	Eating fish is good. It helps your brain.	After researching various **nutrition sources**, I concluded **that** incorporating **omega-3 fatty acids** and **antioxidants** into our diet can significantly **ameliorate cognitive decline** in elderly individuals.

Table 1: Examples illustrating dimensions of coherence, cohesion, and complexity. Blue and orange illustrate distinct aspects of each dimension. In the context of coherence, the blue and orange texts signify different elements that maintain thematic consistency throughout the text. For cohesion, the blue text indicates connectors that link sentences together, while the orange text refers to references to previously mentioned entities. Within complexity, the blue text represents lexical sophistication, whereas the orange text denotes the complexity of sentence structure.

word embeddings and a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) for text scoring purposes. Similarly, Wu et al. (2023) conducts fine-grained annotations on 501 Chinese essays and achieves comparable scoring performance to ChatGPT-3.5 through training based on RoBERTa (Liu et al., 2019). However, these approaches suffer from limited generalizability, being applicable only within the confines of labeled domains.

Other works leverage unsupervised methods to automatically construct data for training purposes. UNION (Guan and Huang, 2020) is trained to differentiate between human-written stories and negative samples. Ru et al. (2023) explores implicit discourse relations with a latent discourse sense, showcasing strong performance.

Furthermore, some studies utilize pre-trained language models to assess text quality without additional training. Shrivastava et al. (2018) evaluates textual coherence by modeling the uncertainty of topics within paragraphs and their interrelations, thus scoring texts. BARTScore (Yuan et al., 2021) and GPTScore (Fu et al., 2023) employ the weighted average of the model's output conditional probabilities as a metric, facilitating multifaceted evaluation across a broad range of generative tasks.

Our work measures the quality of long texts from multiple dimensions, introducing metrics that are task-agnostic and do not necessitate additional training.

3 Method

Long texts, characterized by their extended contexts and abundant information, pose distinct challenges in maintaining textual integrity and quality. We systematically measure the quality of long texts through three dimensions: coherence, cohesion, and complexity. Each dimension is accompanied by corresponding quantitative metrics, allowing for an effective measurement of long text quality.

3.1 Coherence, Cohesion and Complexity

In accordance with linguistic fundamentals, we systematically assess the quality of long texts through the following three dimensions.

Coherence refers to the consistency and clarity of the text as a whole. A coherent text maintains thematic unity throughout its parts, with logical connections between the different sections.

Cohesion measures the degree of tight connection between two sentences or sections of the text, reflected in the use of connectives, pronouns, synonyms, and hypernyms/hyponyms.

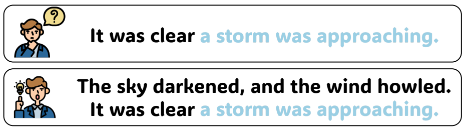
Complexity assesses the level of linguistic sophistication in the use of language in the text. This can be gauged through the richness and diversity of vocabulary, as well as the complexity of sentence structures.

To better elucidate these dimensions, we provide examples in Table 1 that illustrate both high and low levels of these dimensions. Key terms that exemplify specific features of each dimension are highlighted for emphasis.

3.2 Metric

Inspired by the three dimensions mentioned above, we propose the following metrics to assess the quality of long text $\boldsymbol{t} = \{t_1, t_2, \ldots, t_n\}$, including both statistical and model-based ones, where higher values correlate with more pronounced characteristics of the corresponding dimension.

To measure the coherence of a long text, we evaluate the extent to which prior segments of the text contribute to understanding subsequent segments. A coherent text should make it easier to predict its following content based on its preceding context. For example, when predicting the blue text below, it is easier to make a correct prediction if the preceding text is provided.



It was clear a storm was approaching.

The sky darkened, and the wind howled.
It was clear a storm was approaching.

We evaluate the coherence of long texts by comparing the prediction accuracy with a longer context and the accuracy with a shorter context, as well as the difference. Specifically, with a pre-trained causal language model parameterized by θ, we employ the following three metrics for assessing the coherence of long texts:

$$\text{Coherence}_{\text{acc}_l} = \sum_{i=1}^{\lfloor \frac{n}{w} \rfloor} acc\left(\boldsymbol{y}^i | \boldsymbol{x}_l^i, \theta\right) \,/\, \left\lfloor \frac{n}{w} \right\rfloor, \quad (1)$$

$$\text{Coherence}_{\text{acc}_s} = \sum_{i=1}^{\lfloor \frac{n}{w} \rfloor} acc\left(\boldsymbol{y}^i | \boldsymbol{x}_s^i, \theta\right) \,/\, \left\lfloor \frac{n}{w} \right\rfloor, \quad (2)$$

$$\text{Coherence}_{\text{diff}} = \frac{\sum_{i=1}^{\lfloor \frac{n}{w} \rfloor} \frac{\ell\left(\boldsymbol{y}^i|\boldsymbol{x}_l^i,\theta\right) - \ell\left(\boldsymbol{y}^i|\boldsymbol{x}_s^i,\theta\right)}{\ell\left(\boldsymbol{y}^i|\boldsymbol{x}_l^i,\theta\right)}}{\left\lfloor \frac{n}{w} \right\rfloor}, \quad (3)$$

$$\begin{aligned} \text{where } \boldsymbol{x}_l^i &= \{t_{(i-1)w}, \ldots, t_{(i-\frac{1}{4})w}\}, \\ \boldsymbol{x}_s^i &= \{t_{(i-\frac{1}{2})w}, \ldots, t_{(i-\frac{1}{4})w}\}, \\ \boldsymbol{y}^i &= \{t_{(i-\frac{1}{4})w}, \ldots, t_{iw}\}. \end{aligned} \quad (4)$$

$acc(\boldsymbol{y}|\boldsymbol{x}, \theta)$ and $\ell(\boldsymbol{y}|\boldsymbol{x}, \theta)$ denote the model's average top-1 prediction accuracy and negative log-likelihood loss for generating $\boldsymbol{y}$ given the prompt $\boldsymbol{x}$, parameterized by θ. $\text{Coherence}_{\text{acc}_l}$ and $\text{Coherence}_{\text{acc}_s}$ respectively denote the model's top-1 prediction accuracy with longer and shorter preceding texts, and $\text{Coherence}_{\text{diff}}$ represents the proportional improvement in model performance when using a longer versus a shorter context. We process long texts with a sliding window of size w to avoid exceeding the processing capabilities of the language model, setting w to 4096 in practice.

We quantitatively measure cohesion by analyzing the density of connectives and pronouns in the text and the relationships between adjacent sentences. Connectives play pivotal roles in linking words, sentences, or ideas within sentences and paragraphs. Pronouns, serving as substitutes for nouns or noun phrases, maintain references to specific entities mentioned earlier while avoiding unnecessary repetition.

$$\text{Cohesion}_{\text{conn}} = \frac{N_{\text{conn}}}{n}, \quad (5)$$

$$\text{Cohesion}_{\text{pron}} = \frac{N_{\text{pron}}}{n}, \quad (6)$$

$$\text{Cohesion}_{\text{DMR}} = 1 - \sum_{i=1}^{N} \frac{p(\text{no\_conn}|s_i, s_{i+1})}{N}, \quad (7)$$

where N_{conn} and N_{pron} represent the number of connectives and pronouns in the text, respectively. The comprehensive list of considered connectives and pronouns can be found in the Appendix A. The text $\boldsymbol{t}$ consists of $N+1$ sentences, with s_i denoting the i^{th} sentence in the text. The term $p(\text{no\_conn}|s_i, s_{i+1})$ indicates the probability, as determined using Distributed Marker Representation (DMR) (Ru et al., 2023), that sentences s_i and s_{i+1} are unrelated.[1]

The complexity of the text is assessed from vocabulary and paragraph.

$$\text{Complexity}_{\text{TTR}} = \frac{N_{\text{unique}}}{n}, \quad (8)$$

$$\text{Complexity}_{\text{para}} = \frac{n}{N_{\text{para}}}, \quad (9)$$

where N_{unique} refers to the number of unique tokens in the text, used to calculate the Type-Token Ratio (TTR) (Richards, 1987). N_{para} denotes the number of paragraphs in the text, used to determine the average paragraph length.

4 LongWanjuan

4.1 Dataset Construction

Based on the analysis and metrics discussed previously, we introduce LongWanjuan, a bilingual

[1] The DMR approach is originally considered for English texts only. To process Chinese data, we follow its training methodology and train a Chinese DMR model based on the Wanjuan dataset.

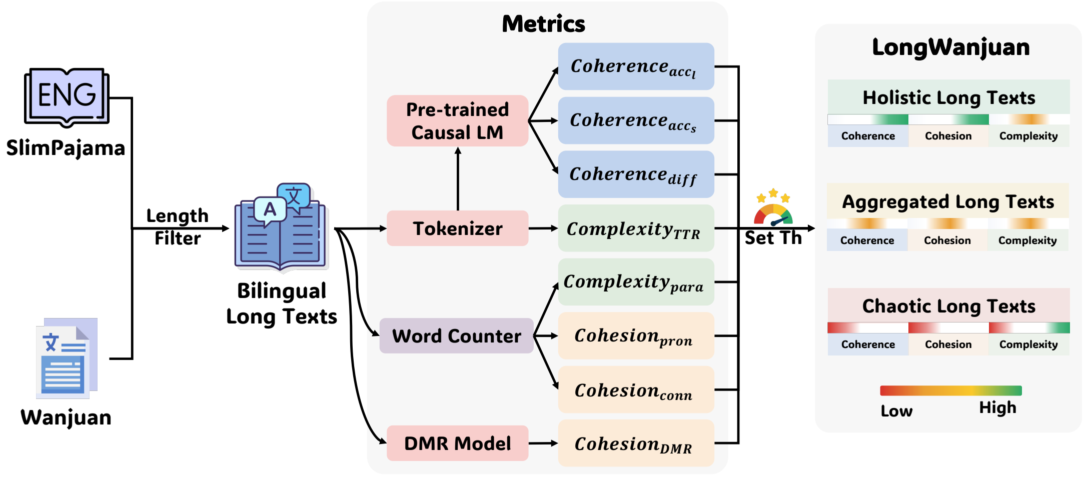


Figure 2: Pipeline for constructing the LongWanjuan dataset.

long-text dataset. The pipeline for constructing our dataset is illustrated in Figure 2.

Given that the majority of the SlimPajama (Soboleva et al., 2023) corpus is in English, we enrich it with Chinese texts from the Wanjuan (He et al., 2023) dataset. Initially, we extract data entries exceeding 32K bytes from both the SlimPajama and Wanjuan datasets, serving as the starting point for our dataset construction.

Subsequently, we evaluate each data entry using the metrics we proposed. Specifically, we first tokenize the data with InternLM2 tokenizer (Team, 2023), thereafter calculating $\text{Complexity}_{\text{TTR}}$. The tokenized results are further processed with InternLM2-7B to obtain coherence scores, including $\text{Coherence}_{\text{acc}_l}$, $\text{Coherence}_{\text{acc}_s}$, and $\text{Coherence}_{\text{diff}}$. We employ NLTK (Bird and Loper, 2004) and LTP (Che et al., 2021) respectively for English and Chinese sentence segmentation. These sentences are then fed into DMR model to derive the $\text{Cohesion}_{\text{DMR}}$ score. The metrics $\text{Cohesion}_{\text{conn}}$, $\text{Cohesion}_{\text{pron}}$ and $\text{Complexity}_{\text{para}}$, are calculated by straightforward word counting.

After scoring each data entry with these metrics, we establish thresholds to categorize the data into holistic long texts, aggregated long texts, and chaotic long texts. During this process, it is necessary only to check whether texts on either side of the threshold belong to different categories. Figure 3 shows the distribution of texts within the C4 domain based on the $\text{Cohesion}_{\text{conn}}$ metric. As illustrated, the texts within different ranges of our proposed metric exhibit distinct characteristics, simplifying the process of threshold determination. For each domain in the dataset, we can extract approximately 30 data samples based on the distribution of this metric and identify the thresholds between different categories of texts. More information on the distribution of text quality across various metrics are shown in Appendix C. In this phase, we initially determine thresholds to segregate holistic long texts. Subsequently, within the remaining texts, we establish thresholds to differentiate chaotic long texts, with the residual texts classified as aggregated long texts.

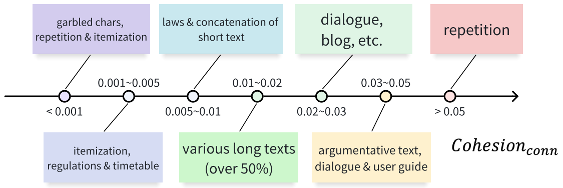

garbled chars, repetition & itemization
laws & concatenation of short text
dialogue, blog, etc.
repetition
0.001~0.005
0.01~0.02
0.03~0.05
< 0.001
0.005~0.01
0.02~0.03
> 0.05
itemization, regulations & timetable
various long texts (over 50%)
argumentative text, dialogue & user guide
$Cohesion_{conn}$

Figure 3: Distribution of texts with different characteristics on the $\text{Cohesion}_{\text{conn}}$ metric in the C4 domain.

Overall, holistic long texts are characterized by high coherence and cohesion, with moderate complexity. Aggregated long texts exhibit lower coherence and cohesion compared to the former. The main feature of chaotic long texts is their complexity, which is anomalously high or low.

4.2 Statistics

The LongWanjuan dataset comprises a total of 160.6B tokens, as tokenized by the InternLM2 tokenizer. Of these, holistic texts constitute 137.6B tokens, accounting for 85.7% of the dataset; aggregated texts make up 21.8 billion tokens, or 13.6%; and chaotic texts comprise 1.2B tokens, representing 0.7%. In this section, we will present statistical information about LongWanjuan, focusing on the

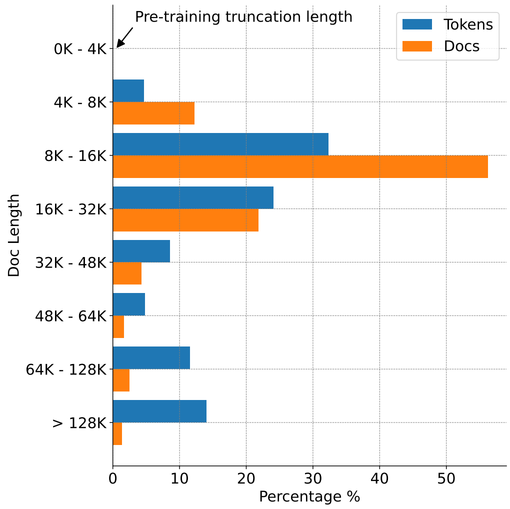


Figure 4: Distribution of token and document counts across different lengths. In LongWanjuan, over 99.9% of the data exceed the truncation length in pre-training.

	Holistic	Aggregated	Chaotic	Total
EN	0.97	0.87	0.81	0.91
ZH	0.97	0.58	0.79	0.80

Table 2: The correlation between manual validation and the classification method we proposed

distribution of domains and lengths. The specific values of token count and document count for each domain are provided in Appendix C.

Length Figure 4 illustrates the distribution of the number of data entries and the number of tokens across different lengths within the LongWanjuan dataset. During pre-training, the training data is generally truncated to a maximum length of 4K tokens, and entries of this length account for less than 0.1% of the dataset in LongWanjuan. In terms of the number of tokens, more than 50% of the data spans lengths between 8K and 32K tokens. Furthermore, over 10% of the data exceeds a length of 128K tokens. With regard to the number of data entries, more than 50% of the documents fall within the 8K to 16K token range. The trend in data entries by length initially increases before decreasing, and due to longer documents containing more tokens, the smallest quantity of tokens is observed in the 48K to 64K range.

5 Experiments

5.1 Manual Validation

	Human Annotators	Annotators v.s. Metrics
EN	0.9095	0.9048
ZH	0.8933	0.6755

Table 3: The kappa score among annotators and that between human validation and classification method.

Complementary to the following training and evaluating results, we conduct human validation by manually annotating the type of 200 long texts from SlimPajama (Soboleva et al., 2023) and Wanjuan (He et al., 2023) and then calculating the classification accuracy. The verification set includes 120 items in English and 80 items in Chinese, covering various domains as well as all three types of long texts in SlimPajama and Wanjuan. The verification results are shown in Table 2 and Table 3.

The quantitative metrics we proposed can effectively distinguish the three types of long texts in SlimPajama and Wanjuan. Specifically, for Chinese, the accuracy of the aggregated long text is relatively low. This is because the 'TextBook' domain in Wanjuan contains a large amount of classical Chinese texts, which have inherent differences compared to modern Chinese texts. On one hand, it is challenging for models and rule-based scoring methods to accurately distinguish between them. On the other hand, there exist difficulties and biases in human annotation. As a result, the relatively lower accuracy is reasonable. Overall, our proposed method can still effectively differentiate the three types of long texts in general Chinese and English language data. In other words, long texts can be classified into these three types from the perspectives of coherence, cohesion, and complexity.

5.2 Setup

We conduct experiments on LLaMA2-7B-4K (Touvron et al., 2023b), LLaMA3-8B-8K (Meta, 2024) and InternLM2-7B (Team, 2023) corresponding to LLMs with and without long context capability respectively. Detailed training hyper-parameters can be found in Appendix D.

For all models, we use a 9:1 ratio of English to Chinese language data. For SlimPajama, we follow the data mixtures used for LLaMA pre-training (Touvron et al., 2023a). Due to the limited amount of Chinese data, we sample data uniformly from Wanjuan. We excluded chaotic texts and up-sample aggregated texts to balance the holistic and aggregated texts as our proposed recipe.

We compare our proposed data-mixing recipe with the following three strategies: 1. Training on long texts from all categories. 2. Training LLM

	EN	ZH	Text	Code	Total
LongChat-v1.5-7B-32K	37.13	14.88	27.63	54.15	33.22
Yi-6B-200K	37.65	15.12	28.04	64.55	35.72
Qwen1.5-7B-128K	42.61	26.33	31.62	**68.45**	39.37
InternLM2-7B	51.61	34.07	40.91	62.86	45.43
ChatGLM3-6B-32K	55.36	**42.43**	45.64	57.10	48.05
LLaMA2-7B with LongWanjuan	33.92	18.94	25.15	62.90	33.10
LLaMA3-8B with LongWanjuan	34.82	20.23	27.55	67.23	35.91
InternLM2-7B with LongWanjuan	**56.64**	39.31	**46.26**	65.26	**50.26**

Table 4: Comparison between our proposed training strategy with other open-sourced LLMs on LongBench. The terms HOL, AGG, and CHA respectively denote holistic texts, aggregated texts, and chaotic texts.

	EN	Δ	ZH	Δ	Text	Δ
LLaMA2-7B-4K	28.55		13.62		21.41	
HOL. + AGG. + CHA.	32.86	+15.11%	17.18	+26.20%	24.30	+13.46%
HOL.	33.17	+16.20%	18.44	+35.44%	24.63	+15.02%
HOL. + AGG.	33.66	+17.91%	17.14	+25.88%	24.99	+16.70%
HOL. + Upsampling AGG.	**33.92**	+18.80%	**18.94**	+39.09%	**25.15**	+17.45%
LLaMA3-8B-8K	33.16		18.86		25.47	
HOL. + AGG. + CHA.	33.54	+1.15%	21.12	+12.02%	26.51	+4.09%
HOL.	33.50	+1.03%	21.71	+15.16%	26.61	+4.46%
HOL. + AGG.	34.00	+2.51%	**23.02**	+22.08%	**27.55**	+8.17%
HOL. + Upsampling AGG.	**34.82**	+5.00%	20.23	+7.30%	**27.55**	+8.17%
InternLM2-7B	51.61		34.07		40.91	
HOL. + AGG. + CHA.	55.03	+6.63%	36.63	+7.52%	44.49	+8.74%
HOL.	55.12	+6.81%	36.97	+8.51%	44.61	+9.04%
HOL. + AGG.	55.54	+7.62%	37.36	+9.67%	44.79	+9.46%
HOL. + Upsampling AGG.	**56.64**	+9.76%	**39.31**	+15.38%	**46.26**	+13.07%

Table 5: Comparison of different training strategies data on LongBench. We also report relative improvements over the pre-trained LLMs in the same way as LLaMA2Long (Xiong et al., 2023). The terms HOL, AGG, and CHA respectively denote holistic texts, aggregated texts, and chaotic texts.

with only the holistic long texts. 3. Excluding chaotic texts and employing holistic and aggregated texts for training.

5.3 Main Results

We first compare the training results of LLaMA2-7B, LLaMA3-8B and InternLM2-7B with our data mixing recipe mentioned above on LongWanjuan with other long-context LLMs, such as LongChat-v1.5-7B-32K (Li et al., 2023), Yi-6B-200K (01-ai, 2023), Qwen1.5-7B-128K (Bai et al., 2023a) and ChatGLM3-6B-32K (Zeng et al., 2023), on LongBench (Bai et al., 2023b), a widely accepted benchmark dataset for long-context LLM. LongBench includes different languages (Chinese and English) and application areas (such as single-doc QA, multi-doc QA, summarization, few-shot learning tasks, synthetic tasks, and code completion) to provide a comprehensive evaluation of the language model's capabilities in handling long contexts. During the evaluation, we limit the maximum input length to 4K tokens for pre-trained LLaMA2-7B-4K, 8K tokens for pre-trained LLaMA3-8B-8K, and 32K tokens for other models. We apply the truncation from the middle used in LongBench.

The results are shown in Table 4, and detailed scores for each subtask can be found in the Appendix F. Despite the strong long-text capabilities of InternLM2-7B, continuing training on LongWanjuan using our recipe leads to performance improvements across all domains. Moreover, we surpassed ChatGLM3-6B-32K overall, achieving a new state-of-the-art performance on LongBench.

5.4 Analysis

Then we compare the training results of LLaMA2-7B, LLaMA3-8B and InternLM2-7B with the three strategies mentioned above. The results are shown in Table 5, and detailed scores for each subtask can be found in Appendix F. Since our work mainly focuses on the quality of long text, we do not emphasize the improvement in code-related abilities. We observed that training solely on holistic texts yielded only marginal improvements compared to

	Single-doc	Multi-doc	Sum	Few-shot	Synthetic
LLaMA2-7B-4K	18.43	11.50	15.24	52.36	**5.34**
HOL. + AGG. + CHA.	**23.71**	12.54	17.32	59.23	3.45
HOL.	23.57	**12.87**	19.43	57.79	4.38
HOL. + AGG.	22.35	12.38	**20.42**	59.68	4.96
HOL. + Upsampling AGG.	22.56	12.74	19.97	**61.14**	3.86
LLaMA3-8B-8K	23.56	13.74	20.97	62.14	4.86
HOL. + AGG. + CHA.	24.56	14.74	21.97	63.14	5.86
HOL.	25.56	15.74	22.97	64.14	6.86
HOL. + AGG.	26.56	16.74	23.97	65.14	7.86
HOL. + Upsampling AGG.	**27.56**	**17.74**	**24.97**	**66.14**	**8.86**
InternLM2-7B	43.50	37.10	23.70	59.95	40.33
HOL. + AGG. + CHA.	42.05	39.96	23.73	61.43	58.67
HOL.	40.46	**40.83**	24.03	62.07	59.00
HOL. + AGG.	42.63	40.35	24.66	61.83	57.50
HOL. + Upsampling AGG.	**44.20**	40.15	**25.28**	**62.70**	**63.05**

Table 6: Comparison of different training strategies data on the major task categories in LongBench. The terms HOL, AGG, and CHA respectively denote holistic texts, aggregated texts, and chaotic texts.

using data from all categories without any filtering. Incorporating aggregated texts can lead to a significant enhancement and achieve optimal performance among these strategies, especially when the ratio of aggregated texts is upsampled.

We analyze the performance of these data mixing strategies across different tasks in Table 6. For LLaMA2-7B, the removal of chaotic texts results in improvements across multi-doc QA, summarization, few-shot learning tasks, and synthetic tasks. Additionally, incorporating aggregated texts alongside training solely on holistic texts enhances performance on these tasks. Although our proposed recipe excels primarily in few-shot learning tasks, it demonstrates overall superior performance. Regarding InternLM2-7B, our proposed recipe achieves optimal performance across all tasks except for multi-doc QA. We attribute the differing performances between the two models to the relatively lower proportion of Chinese in LLaMA2-7B's pretraining corpus compared to our continued training with a 10% Chinese ratio. Despite this distinction, our recipe yields the best overall performance on both these models. When it comes to LLaMA3-8B, our proposed method acquires the optimal performance across all task types.

To validate the generality of our LongWanjuan, we compare the performance of models trained with different data mixing strategies on another commonly used long-context evaluation dataset, L-Eval(An et al., 2023). The results are shown in Table 11 in Appendix E. It can be found that the models trained on the filtered text significantly outperformed those trained on the unfiltered text. Due to the lack of multi-doc-related tasks in L-Eval, the addition of aggregated texts had a limited impact on model performance. Nevertheless, fine-tuning with the holistic and aggregated text still performs best on average. Our proposed data mixing also achieves the best results for InternLM2-7B.

We also evaluate the performance of models fine-tuned on long texts across multiple short tasks with a length of less than 2K tokens. Our findings indicate that the average performance fluctuation remains within 1.5 percentage points. Furthermore, incorporating aggregated texts proves to be effective in enhancing performance on short tasks. For detailed performance metrics and benchmark test results, please refer to the Appendix G.

6 Conclusion

We try to systematically analyze the quality of long texts from three linguistic dimensions: coherence, cohesion, and complexity. Inspired by these dimensions, we develop a series of metrics based on statistics and pre-trained models to quantitatively assess the quality of long texts. Utilizing SlimPajama and Wanjuan, we construct the LongWanjuan dataset and categorize texts into three types: holistic, aggregated, and chaotic texts, according to our proposed metrics. We introduce a data mixture recipe based on the LongWanjuan dataset to address the issue of the imbalance between holistic long texts and aggregated long texts, achieving state-of-the-art performance on the LongBench benchmark. Our experimental analysis further validates the effectiveness of the proposed recipe.

Limitations

We utilize SlimPajama and Wanjuan to construct LongWanjuan, with the Chinese data still remaining relatively limited. Based on the scalability and generalizability of our approach, additional Chinese datasets and datasets from other languages can be incorporated on top of deduplication. We alleviate the imbalance between the quantities of holistic and aggregated texts by upsampling aggregated texts. However, we did not attempt to provide an optimal ratio, leaving this for future work.

Ethics Statement

LongWanjuan is constructed based on Wanjuan (under the CC BY 4.0 license) and SlimPajama (under the Apache 2.0 license), both of which permit open and free usage. We plan to open-source LongWanjuan under the CC BY 4.0 license.

Throughout the dataset construction process, there are 3 annotators involved, all of whom are authors. The annotators are all native Chinese speaker and proficient in reading and understanding English. They consent to contribute their efforts to building LongWanjuan.

References

01-ai. 2023. Yi: Building the next generation of open-source and bilingual llms.

Amro Abbas, Kushal Tirumala, Daniel Simig, Surya Ganguli, and Ari S. Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *CoRR*, abs/2303.09540.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023b. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

Stella Biderman, USVSN Sai Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. 2023. Emergent and predictable memorization in large language models. *CoRR*, abs/2304.11158.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Patricia L Carrell. 1982. Cohesion is not coherence. *TESOL quarterly*, 16(4):479–488.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An open-source neural language technology platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595.

Woon Sang Cho, Pengchuan Zhang, Yizhe Zhang, Xiujun Li, Michel Galley, Chris Brockett, Mengdi Wang, and Jianfeng Gao. Towards coherent and cohesive long-form text generation. *NAACL HLT 2019*, page 1.

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*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4599–4610. Association for Computational Linguistics.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten P. Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathleen S. Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V. Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. 2022. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 5547–5569. PMLR.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *CoRR*, abs/2302.04166.

Jian Guan and Minlie Huang. 2020. UNION: an unreferenced metric for evaluating open-ended story generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9157–9166. Association for Computational Linguistics.

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

Michael Alexander Kirkwood Halliday and Ruqaiya Hasan. 2014. *Cohesion in english*. 9. Routledge.

Conghui He, Zhenjiang Jin, Chao Xu, Jiantao Qiu, Bin Wang, Wei Li, Hang Yan, Jiaqi Wang, and Dahua Lin. 2023. Wanjuan: A comprehensive multimodal dataset for advancing english and chinese large models. *arXiv preprint arXiv:2308.10755*.

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*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Uniform complexity for text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12025–12046.

Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. 2023. How long can context length of open-source llms truly promise? In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Wenbiao Li, Wang Ziyang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7446–7457.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3252. Association for Computational Linguistics.

Tianyang Liu, Canwen Xu, and Julian J. McAuley. 2023a. Repobench: Benchmarking repository-level code auto-completion systems. *CoRR*, abs/2306.03091.

Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. 2023b. Scaling laws of rope-based extrapolation. *CoRR*, abs/2310.05209.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2023. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *CoRR*, abs/2310.07931.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*, abs/2309.04564.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Asap++: Enriching the asap automated essay grading dataset with essay attribute scores. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Gabriele Pallotti. 2015. A simple view of linguistic complexity. *Second Language Research*, 31(1):117–134.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. 2021. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 20596–20607.

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. *CoRR*, abs/2306.01116.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023. Yarn: Efficient context window extension of large language models. *CoRR*, abs/2309.00071.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. 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 M. 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 J. Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *CoRR*, abs/2112.11446.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton-Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2023. Code llama: Open foundation models for code. *CoRR*, abs/2308.12950.

Dongyu Ru, Lin Qiu, Xipeng Qiu, Yue Zhang, and Zheng Zhang. 2023. Distributed marker representation for ambiguous discourse markers and entangled relations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5334–5351. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Muhammad Salman, Armin Haller, and Sergio J Rodríguez Méndez. 2023. Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification. *arXiv preprint arXiv:2304.07774*.

Disha Shrivastava, Abhijit Mishra, and Karthik Sankaranarayanan. 2018. Modeling topical coherence in discourse without supervision. *CoRR*, abs/1809.00410.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Xin Tan, Longyin Zhang, and Guodong Zhou. 2022. Discourse cohesion evaluation for document-level neural machine translation. *arXiv preprint arXiv:2208.09118*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari S. Morcos. 2023. D4: improving LLM pretraining via document de-duplication and diversification. *CoRR*, abs/2308.12284.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Yuan Wang and Minghe Guo. 2014. A short analysis of discourse coherence. *Journal of Language Teaching and Research*, 5(2):460.

Hongyi Wu, Xinshu Shen, Man Lan, Shaoguang Mao, Xiaopeng Bai, and Yuanbin Wu. 2023. A multi-task dataset for assessing discourse coherence in chinese essays: Structure, theme, and logic analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6673–6688.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. Effective long-context scaling of foundation models. *CoRR*, abs/2309.16039.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Retrieval meets long context large language models. *CoRR*, abs/2310.03025.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2023. Data-centric artificial intelligence: A survey. *CoRR*, abs/2303.10158.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. *Recall*, 4(F1):5–19.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir R. Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5905–5921. Association for Computational Linguistics.

A Connectives and Pronouns

The connectives and pronouns utilized in our metric calculations are outlined in Table 7 and Table 8, respectively.

B Distribution of Texts across Metrics

In this section, we report the distribution features with more characteristics, including Cohesion$_{\text{conn}}$, Cohesion$_{\text{pron}}$, Cohesion$_{\text{DMR}}$, Complexity$_{\text{para}}$, in Figure 5 to Figure 11. We take the C4 domain and the ChinaNews domain as an example of English and Chinese texts respectively.

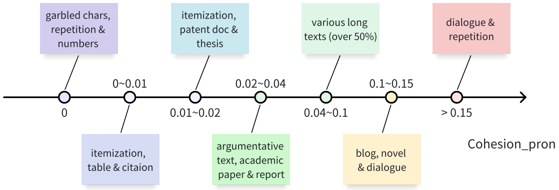

garbled chars, repetition & numbers
itemization, patent doc & thesis
various long texts (over 50%)
dialogue & repetition
0~0.01
0.02~0.04
0.1~0.15
0
0.01~0.02
0.04~0.1
> 0.15
itemization, table & citaion
argumentative text, academic paper & report
blog, novel & dialogue
Cohesion_pron

Figure 5: Distribution of texts with different characteristics on the Cohesion$_{\text{pron}}$ metric in the C4 domain.

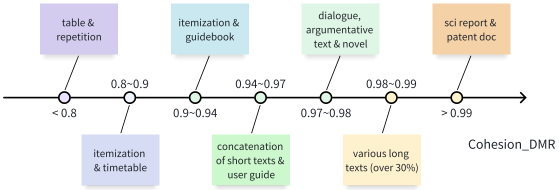


Figure 6: Distribution of texts with different characteristics on the Cohesion$_{\text{DMR}}$ metric in the C4 domain.

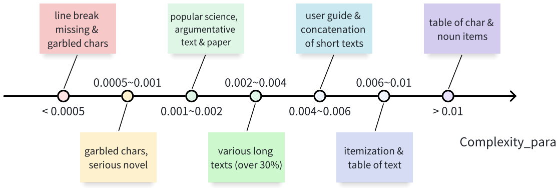

line break missing & garbled chars
popular science, argumentative text & paper
user guide & concatenation of short texts
table of char & noun items
0.0005~0.001
0.002~0.004
0.006~0.01
< 0.0005
0.001~0.002
0.004~0.006
> 0.01
garbled chars, serious novel
various long texts (over 30%)
itemization & table of text
Complexity_para

Figure 7: Distribution of texts with different characteristics on the Complexity$_{\text{para}}$ metric in the C4 domain.

Conn. in English	’but ’, ’whereas’, ’however’, ’though’, ’yet’, ’nevertheless’, ’still’, ’despite’, ’nonetheless’, ’notwithstanding’, ’regardless of’, ’in spite of’, ’apart from’, ’in any case’, ’in any event’, ’supposedly’, ’provided’, ’otherwise’, ’unless’, ’once’, ’as long as’, ’because’, ’so ’, ’since’, ’thus’, ’therefore’, ’as a result’, ’accordingly’, ’thereafter’, ’thereby’, ’hence’, ’given’, ’due to’, ’owing to’, ’on account of’, ’in light of’, ’as a matter of fact’, ’in other words’, ’alternatively,’, ’alternately,’, ’optionally,’, ’namely,’, ’that is to say’, ’in contrast’, ’on the contrary’, ’in turn’, ’by contrast’, ’conversely,’, ’by comparison’, ’for example’, ’for instance’, ’typically,’, ’specifically,’, ’especially,’, ’particularly,’, ’in particular’, ’until’, ’while’, ’when’, ’recently,’, ’presently,’, ’currently,’, ’in the meantime’, ’previously,’, ’initially,’, ’originally,’, ’subsequently,’, ’later’, ’consequently,’, ’finally,’, ’ultimately,’, ’eventually,’, ’in the end’, ’lately,’, ’lastly,’, ’firstly,’, ’secondly,’, ’thirdly,’, ’next’, ’on one hand’, ’on the other hand’, ’moreover’, ’in addition’, ’additionally,’, ’besides’, ’furthermore’, ’in sum’, ’in summary’, ’overall’, ’in short’, ’in conclusion’, ’in brief’, ’in detail’, ’personally,’, ’luckily,’, ’thankfully,’, ’fortunately,’, ’hopefully,’, ’preferably,’, ’surprisingly,’, ’ironically,’, ’amazingly,’, ’oddly,’, ’sadly,’, ’historically,’, ’traditionally,’, ’theoretically,’, ’practically,’, ’realistically,’, ’actually,’, ’generally,’, ’ideally,’, ’technically,’, ’honestly,’, ’frankly,’, ’basically,’, ’admittedly,’, ’undoubtedly,’, ’importantly,’, ’essentially,’, ’naturally,’, ’arguably,’, ’remarkably,’, ’in fact’, ’in essence’, ’in practice’, ’in general’, ’by doing this’.
Conn. in Chinese	’至今为止，’, ’目前’, ’这样一来’, ’详细地’, ’与此同时，’, ’起初’, ’换言之’, ’此刻’, ’鉴于’, ’其中，’, ’例如，’, ’突然’, ’那么，’, ’不久，’, ’并且’, ’确实，’, ’尽管’, ’而不是’, ’总体上，’, ’第一，’, ’无论’, ’最近’, ’无论如何’, ’简而言之’, ’这里，’, ’有时候，’, ’除非’, ’结果，’, ’然后，’, ’除开’, ’当然，’, ’很快，’, ’但是，’, ’另一方面，’, ’换句话说，’, ’理论上’, ’历史上’, ’虽然’, ’不管’, ’所以，’, ’首先’, ’而且’, ’而’, ’由于’, ’第三，’, ’可是，’, ’但’, ’由此可见，’, ’而是’, ’最初，’, ’最终，’, ’后来，’, ’即使’, ’只有这样，’, ’但事实上，’, ’相反’, ’总的来说，’, ’只是’, ’取决于’, ’这时，’, ’用来’, ’以便’, ’基本上，’, ’不料’, ’就像’, ’接下来’, ’老实说’, ’相比之下，’, ’本质上’, ’否则，’, ’从某种意义上’, ’之前’, ’当时’, ’以前’, ’以至于’, ’特别是’, ’尤其是’, ’实际上，’, ’只要’, ’理想情况’, ’或者，’, ’不仅如此，’, ’幸运’, ’事实上，’, ’然而，’, ’一方面，’, ’比如，’, ’通常’, ’原因是’, ’从长远来看’, ’此后’, ’其次’, ’渐渐地，’, ’直到’, ’不论’, ’大多数情况下’, ’之后，’, ’显然’, ’也就是说，’, ’以及’, ’随后，’, ’没想到’, ’不过，’, ’除此之外’, ’无疑’, ’第二，’, ’反过来，’, ’若是’, ’以上就是’, ’也许’, ’假如’, ’可’, ’如果’, ’一如既往’, ’结果就是’, ’通过这样’, ’类似地，’, ’一般来说，’, ’除了’, ’据说’, ’另外，’, ’同样地’, ’反之，’, ’总之，’, ’进一步’, ’可以说’, ’于是，’, ’最后，’, ’既然’, ’尽管如此，’, ’这意味着’, ’同时，’, ’因此，’, ’某种程度上’, ’综上，’, ’随着’, ’此外，’, ’即便如此’, ’有时，’, ’同样，’.

Table 7: The connectives we use to calculate $\text{Cohesion}_{\text{conn}}$. These words and phrases are collected from the list of connective words in Ru et al. (2023).

Pron. in English	’one’, ’ones’, ’i’, ’me’, ’my’, ’mine’, ’myself’, ’you’, ’your’, ’yours’, ’yourself’, ’he’, ’him’, ’his’, ’himself’, ’she’, ’her’, ’hers’, ’herself’, ’it’, ’its’, ’itself’, ’we’, ’us’, ’our’, ’ours’, ’ourselves’, ’they’, ’them’, ’their’, ’theirs’, ’themselves’, ’this’, ’that’, ’these’, ’those’, ’who’, ’whom’, ’whose’.
Pron. in Chinese	’我’, ’自己’, ’你’, ’他’, ’她’, ’它’, ’这’, ’那’, ’这个’, ’那个’, ’那里’, ’彼此’, ’您’, ’我们’, ’你们’, ’他们’, ’她们’, ’它们’, ’这些’, ’那些’.

Table 8: The pronouns we use to calculate $\text{Cohesion}_{\text{pron}}$.

C Detailed Statistics

Figures 12a and 12b depict the distribution of data across various domains in English and Chinese, respectively, within the LongWanjuan dataset. In these bar graphs, each row is divided into three segments from left to right, representing holistic texts, aggregated texts, and chaotic texts, in that order. In the English data, the CommonCrawl domain predominates, accounting for over 50% of the data. Apart from a significant amount of aggregated texts in the CommonCrawl domain, the majority of data in other domains consists of holistic texts. In the Chinese data, the distribution across different domains is more balanced, with each domain featuring both holistic and aggregated texts. The WebText and Law domains contain a notable number of chaotic texts. Detailed statistical information is available in Table 9 and Table 10, respectively.

Domain	#Docs				#Tokens			
	Holistic	Aggregated	Chaotic	Total	Holistic	Aggregated	Chaotic	Total
CommonCrawl	4740880	638363	36664	5415907	76.5B	9.9B	719.8M	87.2B
C4	632819	88119	2732	723670	7.0B	1.1B	36.6M	8.2B
ArXiv	1045806	3274	287	1049367	25.4B	153.9M	68.3M	25.6B
Book	187396	7369	252	195017	24.2B	893.9M	80.7M	25.1B
Github	377312	56557	0	433869	7.4B	1.3B	0.0M	8.7B
Wikipedia	146469	29745	1883	178097	2.9B	654.4M	97.8M	3.7B
StackExchange	5295	1750	659	7704	60.6M	21.9M	11.3M	93.8M
Total	6856817	786654	48564	7692035	137.6B	13.0B	1.2B	151.8B

Table 9: An overview of the dataset statistics in the English part of LongWanjuan. The number of tokens is calculated with the tokenizer in InternLM2-7B (Team, 2023).

Domain	#Docs				#Tokens			
	Holistic	Aggregated	Chaotic	Total	Holistic	Aggregated	Chaotic	Total
ChinaNews	5211	1331	240	6782	51.3M	15.5M	4.3M	71.1M
Law	24575	5212	1310	31097	276.3M	58.1M	69.4M	403.8M
Patent	44922	2956	682	48560	438.0M	31.6M	9.9M	479.5M
TextBook	4746	693	0	5439	496.0M	119.3M	0.0M	615.3M
WebText	18698	7842	3855	30395	180.6M	93.0M	91.4M	365.1M
Total	98152	18034	6087	122273	1.4B	317.4M	175.1M	1.9B

Table 10: An overview of the dataset statistics in the Chinese part of LongWanjuan. The number of tokens is calculated with the tokenizer in InternLM2-7B (Team, 2023).

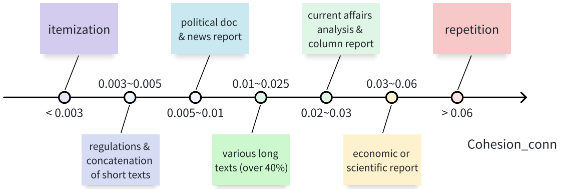

itemization
political doc & news report
current affairs analysis & column report
repetition
< 0.003
0.003~0.005
0.005~0.01
0.01~0.025
0.02~0.03
0.03~0.06
> 0.06
regulations & concatenation of short texts
various long texts (over 40%)
economic or scientific report
Cohesion_conn

Figure 8: Distribution of texts with different characteristics on the Cohesion$_{\text{conn}}$ metric in the ChinaNews domain.

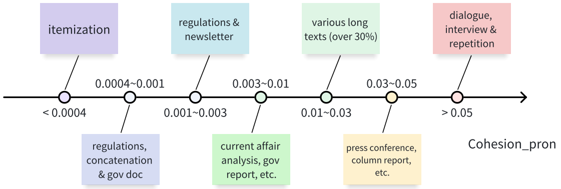

itemization
regulations & newsletter
various long texts (over 30%)
dialogue, interview & repetition
< 0.0004
0.0004~0.001
0.001~0.003
0.003~0.01
0.01~0.03
0.03~0.05
> 0.05
regulations, concatenation & gov doc
current affair analysis, gov report, etc.
press conference, column report, etc.
Cohesion_pron

Figure 9: Distribution of texts with different characteristics on the Cohesion$_{\text{pron}}$ metric in the ChinaNews domain.

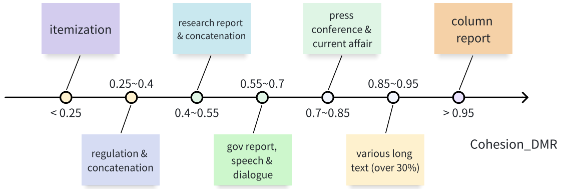

itemization
research report & concatenation
press conference & current affair
column report
< 0.25
0.25~0.4
0.4~0.55
0.55~0.7
0.7~0.85
0.85~0.95
> 0.95
regulation & concatenation
gov report, speech & dialogue
various long text (over 30%)
Cohesion_DMR

Figure 10: Distribution of texts with different characteristics on the Cohesion$_{\text{DMR}}$ metric in the ChinaNews domain.

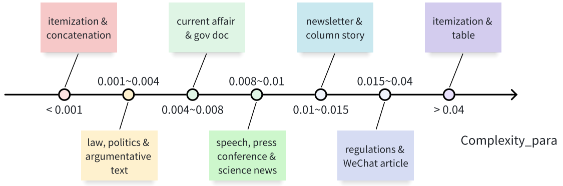
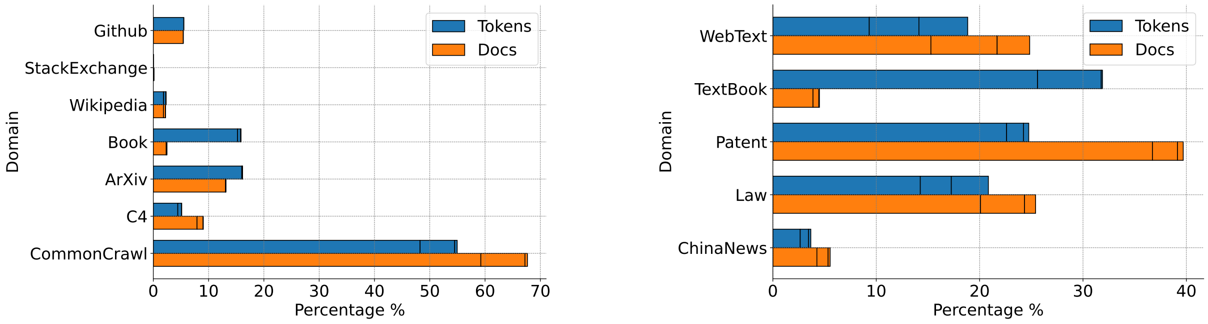

itemization & concatenation
current affair & gov doc
newsletter & column story
itemization & table
< 0.001
0.001~0.004
0.004~0.008
0.008~0.01
0.01~0.015
0.015~0.04
> 0.04
law, politics & argumentative text
speech, press conference & science news
regulations & WeChat article
Complexity_para

Figure 11: Distribution of texts with different characteristics on the Complexity$_{\text{para}}$ metric in the ChinaNews domain.

D Hyper-parameters

We use 64 A100 GPUs and adopt ZeRO3 strategies (Rajbhandari et al., 2020) to tune a 7B model. We use AdamW (Loshchilov and Hutter, 2017) with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. We set the learning rate to 3×10^{-5} with a cosine learning rate schedule with a 20-step warmup. We set the max gradient norm to 1 and the weight decay to zero.

We fine-tune both LLaMA2-7B-4K and InternLM2-7B with 5B tokens using the next token prediction objective. We set the global batch size to 2M tokens, with a max length of 32K tokens. Specifically, for the fine-tuning of LLaMA2-7B to achieve context over 32K tokens, we adjust the base of the rotation angle in RoPE (Su et al., 2024) to 500000 based on LLaMA2Long (Xiong et al., 2023) and ScalingRoPE (Liu et al., 2023b).



(a) Distribution of data from SlimPajama.

(b) Distribution of data from Wanjuan.

Figure 12: Distribution of token and document counts across different domains. Each bar is divided from left to right into three parts: holistic, aggregated, and chaotic texts.

	TOEFL	QuALITY	Coursera	SF	MD	NQ	FQA	CUAD	NQA	Avg.
LLaMA2-7B-4K	37.55	27.72	15.12	4.80	1.12	6.17	8.23	6.11	3.41	12.25
HOL. + AGG. + CHA.	45.72	20.30	23.84	4.35	2.33	5.27	9.90	0.61	5.40	13.08
HOL.	38.66	26.73	21.51	4.25	3.45	6.51	9.30	0.71	6.10	13.02
HOL. + AGG.	48.33	25.25	22.67	4.83	1.01	7.63	9.04	1.42	5.79	**14.00**
HOL. + U. AGG.	40.15	23.27	19.19	4.13	1.80	6.35	9.66	1.51	5.18	12.36
LLaMA3-8B-8K	79.93	54.46	22.67	2.10	8.29	0.00	1.13	2.18	0.44	19.02
HOL. + AGG. + CHA.	72.86	51.49	22.67	6.47	18.41	1.59	3.76	3.65	1.72	20.29
HOL.	71.38	44.55	35.47	7.33	13.26	1.07	6.20	5.46	1.87	20.73
HOL. + AGG.	76.95	51.98	23.26	8.34	17.65	0.41	7.30	5.29	2.12	**21.48**
HOL. + U. AGG.	73.61	49.50	17.44	7.41	16.66	0.35	9.07	4.45	1.41	19.99
InternLM2-7B-200K	83.64	74.26	41.86	2.94	17.75	0.00	0.00	2.31	0.08	24.76
HOL. + AGG. + CHA.	84.76	71.29	52.91	13.45	17.95	1.83	7.35	5.29	1.04	28.43
HOL.	83.64	69.31	54.65	12.33	17.97	2.78	7.28	4.57	1.20	28.19
HOL. + AGG.	82.90	71.78	54.07	13.06	16.62	3.04	6.57	4.85	0.91	28.20
HOL. + U. AGG.	82.53	72.28	55.81	13.00	18.66	3.61	8.74	4.91	1.09	**28.96**

Table 11: Comparison of different training strategies data on subtasks in L-Eval, including TOEFL, QuALITY, Coursera, SFictionQA(SF), MultiDoc2Dial(MD), NQ, LongFQA(FQA), CUAD, NarrativeQA(NQA)

E Performance on L-Eval

The results on LongBench of all the models we tested are shown in Table 11. Since we focus on the performance of the pre-trained model and the quality of the pre-trained data, we omitted certain instruction-following oriented tasks in L-Eval in our experiments.

F Detailed Results on LongBench

Detailed results on LongBench of all the models we tested are shown in Table 12, Table 13 and Table 14.

G Performance on Short Tasks

To verify that the LLM trained on long text in our proposed strategies can still achieve good performance on short-text tasks, we also evaluate our fine-tuned LLaMA2-7B and InternLM2-7B with a maximum input context of 2K tokens on short tasks, including ARC-easy/challenge (Clark et al., 2018), Hellaswag (Zellers et al., 2019), Winogrande (Sakaguchi et al., 2021), TruthfulQA (Lin et al., 2022), SuperGLUE (Wang et al., 2019), GSM8K (Cobbe et al., 2021) and MMLU (Hendrycks et al., 2020). The results are shown in Table 15.

	Narrative QA	**Qasper**	**MF_en**	**MF_zh**	**Hotpot QA**	**2Wikim QA**	**Musique**	**Dureader**
LLaMA2-7B-4K	16.86	15.35	23.78	19.08	7.85	10.54	4.27	23.34
HOL. + AGG. + CHA.	22.61	20.39	30.60	22.96	9.34	10.78	6.01	24.01
HOL.	15.36	19.12	35.04	27.64	9.74	10.83	6.00	24.89
HOL. + AGG.	19.15	19.68	29.60	22.78	10.36	10.49	5.47	23.19
HOL. + Upsampling AGG.	16.93	20.16	26.43	27.68	9.63	10.82	6.75	23.77
LLaMA3-8B-8K	18.02	18.78	20.90	20.48	9.11	10.41	6.98	24.45
HOL. + AGG. + CHA.	2.52	19.56	23.57	20.71	10.00	11.50	6.92	19.81
HOL.	6.59	20.17	22.17	20.61	10.56	11.58	6.78	19.91
HOL. + AGG.	6.74	18.40	24.84	22.26	9.80	12.88	6.47	20.28
HOL. + Upsampling AGG.	3.14	18.82	26.05	20.65	8.89	11.07	6.32	19.00
InternLM2-7B	24.02	41.97	47.95	61.16	52.98	37.89	28.02	29.52
HOL. + AGG. + CHA.	26.86	39.95	41.28	59.90	54.76	43.03	31.04	31.00
HOL.	22.52	40.46	39.99	58.76	54.77	45.07	32.28	31.18
HOL. + AGG.	27.25	40.29	42.92	60.14	53.75	44.53	30.87	32.25
HOL. + Upsampling AGG.	29.93	39.62	50.17	58.57	53.68	42.31	32.14	32.46
LongChat-v1.5-7B-32K	16.90	27.70	41.40	29.10	31.50	20.60	9.70	19.50
Yi-6B-200K	12.36	26.41	36.78	22.36	46.57	40.38	25.78	14.73
Qwen1.5-7B-128K	22.59	23.93	46.99	59.39	20.81	16.36	15.99	36.90
ChatGLM3-6B-32K	9.21	43.07	50.86	60.33	55.33	43.73	38.94	41.89

Table 12: Results on single-doc and multi-doc QA subtasks in Longbench including NarrativeQA, Qasper, MultiField_en (MF_en), MultiField_zh (MF_zh), HotpotQA, 2WikimQA, Musique, and Dureader.

	Gov Report	**QMSum**	**MultiNews**	**VCSum**	**TREC**	**Trivia QA**	**SAM Sum**	**LSHT**
LLaMA2-7B-4K	27.09	20.63	3.21	10.02	68.00	89.09	32.09	20.25
HOL. + AGG. + CHA.	29.54	21.75	6.61	11.37	70.00	86.75	39.15	41.00
HOL.	28.66	21.35	16.34	11.36	69.00	88.44	32.71	41.00
HOL. + AGG.	30.72	21.58	18.26	11.11	71.00	88.36	39.36	40.00
HOL. + Upsampling AGG.	28.87	22.14	16.46	12.42	71.50	88.78	39.78	44.50
LLaMA3-8B-8K	31.31	23.00	12.50	7.09	72.00	89.74	44.89	39.50
HOL. + AGG. + CHA.	30.61	22.57	24.78	9.31	74.50	89.03	46.28	41.00
HOL.	30.35	22.84	25.48	8.52	75.00	89.46	44.28	43.00
HOL. + AGG.	30.22	23.24	25.36	5.84	73.00	89.15	44.26	40.50
HOL. + Upsampling AGG.	30.47	22.94	24.73	3.89	70.00	89.45	44.45	40.00
InternLM2-7B	30.02	23.09	26.46	15.23	75.50	92.36	30.94	41.00
HOL. + AGG. + CHA.	33.69	25.03	27.14	9.05	76.00	89.41	37.99	42.33
HOL.	33.68	25.29	27.04	10.12	77.00	89.17	38.85	43.25
HOL. + AGG.	33.49	25.64	27.54	11.95	77.00	89.07	37.43	43.83
HOL. + Upsampling AGG.	32.96	25.49	27.84	14.81	77.00	91.29	41.00	41.50
LongChat-v1.5-7B-32K	30.80	22.70	26.40	9.90	63.50	82.30	34.20	23.20
Yi-6B-200K	29.34	20.65	27.14	8.14	73.50	86.94	9.85	37.50
Qwen1.5-7B-128K	31.17	25.39	26.00	16.17	73.00	89.39	42.51	38.50
ChatGLM3-6B-32K	35.99	24.68	27.44	15.83	79.00	87.39	17.72	42.00

Table 13: Results on summarization and few-shot learning subtasks in Longbench including GovReport, QMSum, MultiNews, VCSum, TREC, TriviaQA, SAMSum, and LSHT.

	PC	PR_en	PR_zh	LCC	Repobench-p
LLaMA2-7B-4K	1.50	5.52	9.00	68.22	62.25
HOL. + AGG. + CHA.	2.05	4.55	3.75	65.17	60.91
HOL.	2.00	5.38	5.75	65.97	61.33
HOL. + AGG.	1.50	7.62	5.75	65.10	60.52
HOL. + Upsampling AGG.	2.50	3.82	5.25	65.93	59.86
LLaMA3-8B-8K	4.21	7.85	21.61	72.36	67.14
HOL. + AGG. + CHA.	1.82	12.64	35.90	67.57	67.25
HOL.	0.72	8.20	38.24	69.98	67.07
HOL. + AGG.	1.09	19.32	49.23	68.54	66.62
HOL. + Upsampling AGG.	1.05	44.37	37.85	67.78	66.68
InternLM2-7B	7.00	56.50	57.50	63.90	61.81
HOL. + AGG. + CHA.	2.00	96.50	77.50	69.96	64.58
HOL.	0.00	98.50	78.50	69.42	65.39
HOL. + AGG.	0.50	96.00	76.00	69.13	65.06
HOL. + Upsampling AGG.	3.14	97.50	88.50	66.80	63.71
LongChat-v1.5-7B-32K	1.00	30.50	7.60	53.00	55.30
Yi-6B-200K	2.50	6.00	7.97	66.10	63.00
Qwen1.5-7B-128K	3.00	9.50	7.00	71.80	65.10
ChatGLM3-6B-32K	2.00	98.50	94.50	60.07	54.12

Table 14: Results on synthetic and code subtasks in Longbench including PassageCount (PC), PassageRetrieval_en (PR_en), PassageRetrieval_zh (PR_zh), LCC and Repobench-p.

	GSM8K	ARC-e	ARC-c	HS	WG	TQA	SG	MMLU	Average
LLaMA2-7B-4K	16.30	52.73	36.95	69.24	61.25	35.09	50.43	46.78	46.10
HOL. + AGG. + CHA.	16.45	53.09	34.24	65.11	61.01	36.11	51.25	44.13	45.17
HOL.	15.54	53.09	33.90	65.46	61.40	34.80	51.40	42.71	44.79
HOL. + AGG.	16.76	54.67	35.93	65.90	61.01	36.40	50.60	44.74	45.75
HOL. + Upsampling AGG.	17.13	53.97	33.22	65.86	60.30	36.26	49.50	44.49	45.09
LLaMA3-8B-8K	49.05	66.49	41.69	72.81	71.51	35.38	51.97	66.09	56.87
HOL. + AGG. + CHA.	41.47	64.73	38.98	72.16	67.32	34.36	48.93	61.97	53.74
HOL.	45.41	66.31	40.34	72.03	68.19	34.65	50.72	61.68	54.92
HOL. + AGG.	43.52	67.37	39.32	71.47	68.03	34.06	48.35	61.63	54.22
HOL. + Upsampling AGG.	46.10	67.20	39.66	72.07	70.09	35.23	49.49	61.87	55.21
InternLM2-7B	69.83	51.50	42.37	54.87	77.35	39.62	78.83	65.60	60.00
HOL. + AGG. + CHA.	69.67	58.38	41.69	64.46	78.93	37.43	78.43	64.45	61.68
HOL.	70.20	50.26	42.37	56.87	77.90	38.30	79.01	64.75	59.96
HOL. + AGG.	70.43	55.56	40.34	61.64	77.43	37.57	78.85	64.11	60.74
HOL. + Upsampling AGG.	68.99	57.14	41.69	65.46	78.61	38.30	79.20	64.11	61.69

Table 15: Results on 0-shot ARC-easy/challenge, Hellaswag (HS), Winogrande (WG), TruthfulQA (TQA), SuperGLUE (SG), 4-shot GSM8K and 5-shot MMLU.