

Ultra-FineWeb: Efficient Data Filtering and Verification for High-Quality LLM Training Data

Yudong Wang¹, Zixuan Fu^{2,3}, Jie Cai¹, Peijun Tang¹, Hongya Lyu¹, Yewei Fang¹, Zhi Zheng¹, Jie Zhou¹, Guoyang Zeng¹, Chaojun Xiao², Xu Han², Zhiyuan Liu²

¹ModelBest Inc. ²Tsinghua University ³Soochow University

wangyudong@modelbest.cn xcjthu@gmail.com

{han-xu,liuzy}@tsinghua.edu.cn



<https://huggingface.co/datasets/openbmb/UltraFineWeb>

Abstract

Data quality has become a key factor in enhancing model performance with the rapid development of large language models (LLMs). Model-driven data filtering has increasingly become a primary approach for acquiring high-quality data. However, it still faces two main challenges: (1) the lack of an efficient data verification strategy makes it difficult to provide timely feedback on data quality; and (2) the selection of seed data for training classifiers lacks clear criteria and relies heavily on human expertise, introducing a degree of subjectivity. To address the first challenge, we introduce an efficient verification strategy that enables rapid evaluation of the impact of data on LLM training with minimal computational cost. To tackle the second challenge, we build upon the assumption that high-quality seed data is beneficial for LLM training, and by integrating the proposed verification strategy, we optimize the selection of positive and negative samples and propose an efficient data filtering pipeline. This pipeline not only improves filtering efficiency, classifier quality, and robustness, but also significantly reduces experimental and inference costs. In addition, to efficiently filter high-quality data, we employ a lightweight classifier based on *fastText*, and successfully apply the filtering pipeline to two widely-used pre-training corpora, *FineWeb* and *Chinese FineWeb* datasets, resulting in the creation of the higher-quality **Ultra-FineWeb** dataset. Ultra-FineWeb contains approximately 1 trillion English tokens and 120 billion Chinese tokens. Empirical results demonstrate that the LLMs trained on Ultra-FineWeb exhibit significant performance improvements across multiple benchmark tasks, validating the effectiveness of our pipeline in enhancing both data quality and training efficiency.

1 Introduction

In recent years, Large Language Models (LLMs) (Touvron et al., 2023; Han et al., 2021; Ouyang et al., 2022; Hu et al., 2024; Team, 2023; Cai et al., 2024; Bai et al., 2023) have achieved remarkable breakthroughs, demonstrating their powerful capabilities in various fields such as code generation (Song et al., 2024; Guo et al., 2024), logical reasoning (Guo et al., 2025; Lyu et al., 2025), and scientific research (Luo et al., 2025; Zhang et al., 2024). Existing studies have indicated that large-scale high-quality (information-intensive) pre-training data is a key factor in driving the continuous improvement of LLMs' capabilities (Penedo et al., 2024; Li et al., 2024; Gunasekar et al., 2023; Xiao et al., 2024).

To construct information-intensive pretraining corpora, current predominant approaches involve selective filtering of massive and noisy internet data sources (Crawl, 2007). Early approaches usually rely on heuristic filtering using hand-crafted rules (Raffel et al., 2020; Weber et al., 2025; Rae et al., 2021; Wenzek et al., 2019) and deduplication (Lee et al., 2021). With increasing demands for better data quality, these heuristic approaches cannot identify complex content noise and lead to suboptimal LLM performance. Thus, model-driven data filtering, which employs a neural classifier to select high-quality content, has emerged as a better choice (Gunasekar et al., 2023; Shao et al., 2024). Notably, datasets such as FineWeb-edu (Penedo et al., 2024), Chinese FineWeb-edu (Yu et al., 2025), CCI3-HQ (Wang et al., 2024), and DCLM (Li et al., 2024) have

demonstrated the efficacy of this paradigm by incorporating model-based classifiers following preprocessing stages, achieving not only substantial improvements in dataset quality but also measurable enhancements in downstream LLM performance across various benchmark tasks. Nevertheless, existing model-driven filtering methods still suffer from two main challenges: (1) There is a lack of efficient validation to quickly verify the filtering results, typically requiring large-scale training to observe the effect, resulting in high costs and low efficiency. (2) They heavily rely on manually-selected seed data, and the data for training classifiers often depends on human expertise, introducing significant subjectivity.

To address these challenges, we design an efficient data filtering pipeline based on an efficient verification strategy. This verification approach enables impartial seed data selection and facilitates iterative data filtering processes. Specifically, in contrast to conventional approaches that verify data quality by training LLMs from scratch using candidate corpora, our proposed efficient verification strategy leverages a nearly-trained LLM as a foundation. We incorporate candidate corpora during the final training steps and utilize the resulting performance improvement as a metric for assessing data quality. This verification strategy significantly enhances evaluation efficiency while maintaining quality assessment accuracy. Based on our efficient verification strategy, we can impartially select high-quality seed data for classifier training. Building upon the assumption that “high-quality seed data is beneficial for LLM training”, we develop and optimize the strategy for selecting classifier training seeds and recipes, while carefully curating balanced sets of both positive and negative samples to ensure classifier quality and robustness. In addition, to effectively reduce computational cost, we adopt a lightweight classifier based on fastText (Joulin et al., 2016). Compared to LLM-based classifiers (Penedo et al., 2024), the fastText-based classifier demonstrates superior inference efficiency, enabling both the filtering of higher-quality training data for LLMs and significantly accelerating the high-quality data filtration pipeline. We apply the proposed data filtering pipeline to the FineWeb (Penedo et al., 2024) and Chinese FineWeb (Yu et al., 2025) datasets (source data from Chinese FineWeb-edu-v2, which includes IndustryCorpus2 (Shi et al., 2024), MiChao (Liu et al., 2023), WuDuo (BAAI, 2023), SkyPile (Wei et al., 2023), WanJuan (Qiu et al., 2024), ChineseWebText (Chen et al., 2023), TeleChat (He et al., 2024), and CCI3 (Wang et al., 2024)), resulting in the creation of higher-quality *Ultra-FineWeb-en* and *Ultra-FineWeb-zh* datasets, collectively referred to as ***Ultra-FineWeb***. Experimental results show that LLMs trained on ***Ultra-FineWeb*** perform excellently across multiple benchmark tasks, providing empirical validation for the effectiveness of our high-quality data filtering pipeline and its efficiency in reducing computational costs.

Our main contributions are as follows. The datasets and classifier will be released to facilitate the development of LLMs.

- **Efficient Verification Strategy:** We propose a computationally efficient verification strategy that enables rapid evaluation of the impact of data on LLM training performance with minimal computational cost, significantly improving the efficiency of high-quality data filtering experiments.
- **Large-Scale High-Quality Pre-training Datasets:** We design and implement an efficient high-quality data filtering pipeline, applied to the FineWeb and Chinese FineWeb datasets, resulting in the creation of higher-quality *Ultra-FineWeb-en* and *Ultra-FineWeb-zh* datasets, collectively referred to as ***Ultra-FineWeb***. ***Ultra-FineWeb*** contains approximately 1 trillion English tokens and 120 billion Chinese tokens, and can facilitate high-quality LLM training.
- **Lightweight Classifier:** The *Ultra-FineWeb classifier* significantly reduces inference costs, achieving superior performance on extracted text from the same data source, thus validating the effectiveness of our proposed data filtering pipeline in enhancing data quality and training efficiency.

2 Methodology

This section introduces the design and implementation of our efficient, high-quality data filtering pipeline, with the overall workflow illustrated in Figure 1(c). First, in Section 2.2, we present an Efficient Verification Strategy that significantly reduces experimental costs while ensuring the reliability of evaluation results. Subsequently, Section 2.3 outlines our methodology for selecting positive sample seed data for classifier training. Finally, Sections 2.4 and 2.5 introduce classifier training recipes and fastText-based quality filtering, respectively, which together ensure optimal data selection quality and inference efficiency.

2.1 Overall Workflow

The overall workflow of the proposed efficient verification-based high-quality filtering pipeline is illustrated in Figure 1(c). We begin by constructing an initial candidate seed pool and applying our efficient verification strategy to identify high-quality samples that significantly improve training performance. These verified

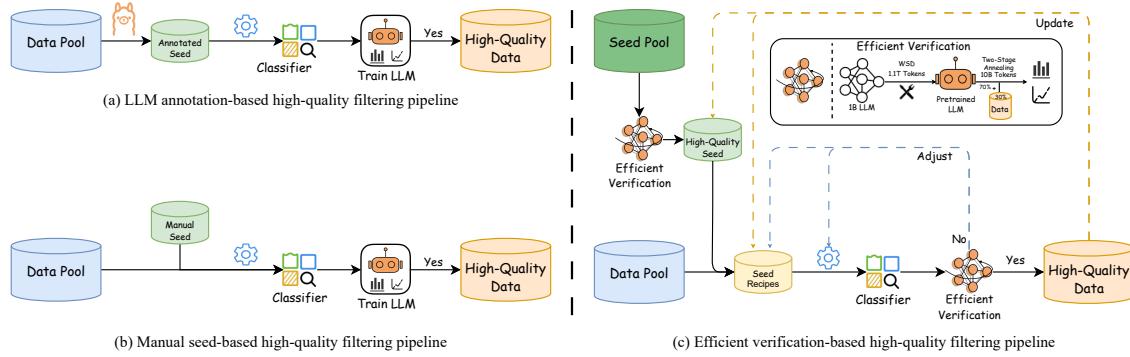


Figure 1: Comparison of High-Quality Data Filtering Pipelines. Traditional model-based data filtering methods (a) and (b) rely on human expertise for seed data selection and lack data quality verification.

samples serve as positive seeds for training a classifier, while negative samples are randomly selected from the raw data pool to create a balanced training set. During the classifier filtering stage, we sample a small subset from the raw data pool and validate the classifier’s selections using our efficient verification strategy to assess its effectiveness. Based on verification results, we iteratively update the high-quality seed pool, adjust the ratio of positive and negative samples, and fine-tune classifier training hyperparameters to optimize the data selection strategy. Only classifiers demonstrating stable and reliable performance in efficient verification are deployed for full-scale data selection and subsequent model training, thereby significantly reducing computational costs while maintaining high data quality.

2.2 Efficient Verification Strategy

Validating the effectiveness of training data typically requires significant computational resources. For instance, training a 1 billion (B) LLM on 100B tokens requires approximately 1,200 H100 GPU hours (equivalent to 64 GPUs running continuously for nearly 19 hours). This computational burden becomes particularly prohibitive when iteratively developing high-quality data classifiers. Moreover, large-scale training validation proves impractical for smaller datasets, as models trained with limited token counts fail to exhibit statistically significant performance differences, with training instability further compromising result reliability. This limitation is evident in our comparative analysis of FineWeb and FineWeb-edu (Penedo et al., 2024). When trained from scratch with 8 billion tokens, FineWeb-edu achieves superior performance on HellaSwag (Zellers et al., 2019), while at 380 billion tokens, FineWeb demonstrates better results across multiple benchmarks, including Winogrande (Sakaguchi et al., 2021), HellaSwag (Zellers et al., 2019), and PIQA (Bisk et al., 2020), highlighting the inconsistency in evaluation outcomes based on training scale¹.

Inspired by Llama 3.1 (Dubey et al., 2024), we design an Efficient Verification Strategy. We begin by training a 1B LLM on 1.1 trillion (T) tokens using a WSD scheduler (Hu et al., 2024) (comprising stable training on 1T tokens, followed by decay training on 0.1T tokens). Based on this pretrained LLM, we then implement a two-stage annealing process with 10B tokens, allocating 30% of the weight to the verification data, while keeping the remaining 70% for the default mixed data ratio. Model details and training hyperparameters can be found in Appendix A. As shown in Table 1, this optimized strategy reduces computational costs from 1,200 to approximately 110 H100 GPU hours (equivalent to less than 3.5 hours on 32 GPUs), significantly reducing training costs and effectively improving the efficiency and iterability of the filtering process, with the two-stage annealing results using the original mixed data ratio as the baseline. This strategy allows for efficient assessment of the impact of verification data across various evaluation dimensions. To validate the reliability of this strategy, we compare the results of training 100B tokens from scratch on the 1B LLM using FineWeb and FineWeb-edu, respectively. As shown in Table 14, the results follow similar trends, with further experimental analysis provided in Appendix B.

2.3 Classifier Training Seeds

The effectiveness of high-quality data classifiers fundamentally depends on the selection of superior positive training samples. As illustrated in Figure 1(a), datasets such as FineWeb-edu (Penedo et al., 2024), Chinese-FineWeb-edu (Yu et al., 2025), and CCI3-HQ (Wang et al., 2024) employ LLM annotation-based frameworks

¹<https://huggingface.co/spaces/HuggingFaceFW/blogpost-fineweb-v1>

Table 1: Comparison of computational costs across different verification strategies on a 1B LLM.

	100B from scratch	380B from scratch	Efficient Verification Strategy
GPU Hours	1,200	4,600	110

to partially label source-consistent data, generating “seed data”. In contrast, Figure 1(b) demonstrates manual seed-based filtering (DCLM’s (Li et al., 2024)) pipeline, which relies on manual curation for positive sample selection, focusing specifically on instruction-formatted data by incorporating samples from OpenHermes 2.5 (OH-2.5) (Teknium, 2023) and high-quality posts from the r/ExplainLikeImFive (ELI5) subreddit.

Although both pipelines demonstrate distinct advantages in selecting positive samples, they are accompanied by inherent limitations. The LLM annotation-based pipeline can effectively filter high-quality samples from source-consistent data, but its performance is constrained by the scoring criteria of the LLM, potentially introducing systematic biases and annotation noise. Furthermore, classifiers trained exclusively on source-consistent data often exhibit limited generalization capabilities and poor robustness. Conversely, manual curation faces significant methodological challenges: the effectiveness of seed data is difficult to assess before classifier training, and its validation relies heavily on the performance of LLMs trained on the filtered data. These constraints lead to high computational costs and reduced adaptability across different tasks.

Based on these considerations, we propose a key assumption: high-quality seed data that enhances LLM performance will yield classifiers capable of identifying similarly beneficial training data. As illustrated in Figure 1 (c), we implement our Efficient Verification Strategy to rapidly evaluate and validate seed data quality within the candidate pool, ensuring the selection of samples that can improve LLM training results. This pipeline not only ensures superior data quality but also optimizes filtration efficiency, thereby generating more reliable positive samples for classifier training. Furthermore, to enhance classifier robustness, we expand negative sample selection beyond source-consistent data. Experimental results further demonstrate that incorporating diverse data sources for negative samples can improve the generalizability of the classifier.

2.4 Classifier Training Recipes

We evaluate a large pool of candidate seed data and ultimately select those with clear effectiveness as positive samples. The positive samples include: (1) LLM-annotated data with scores above 4^{2,3}; (2) instruction-formatted datasets such as OH-2.5 and ELI5; (3) authentic textbook data; (4) LLM-synthesized educational data; and (5) high-quality web content obtained through targeted crawling. For negative samples, we incorporate raw data from diverse sources, including English corpora (FineWeb (Penedo et al., 2024), C4 (Raffel et al., 2020), Dolma (Soldaini et al., 2024), Pile (Gao et al., 2020), and RedPajama (Weber et al., 2025)) and Chinese datasets (IndustryCorpus2 (Shi et al., 2024), MiChao (Liu et al., 2023), WuDao (BAAI, 2023), SkyPile (Wei et al., 2023), WanJuan (Qiu et al., 2024), ChineseWebText (Chen et al., 2023), TeleChat (He et al., 2024), and CCI3 (Wang et al., 2024)) in the initial iteration. To maintain dataset diversity and balance, we implement a uniform distribution strategy, with underrepresented categories undergoing 3-5 rounds of strategic resampling.

Subsequently, we conduct a single iteration of the classifier, utilizing its current predictions as training data for the next round. However, empirical results indicate that the iterative process only contributed meaningfully in the first round, as subsequent updates do not yield further performance improvements and, in some cases, even lead to a decrease in LLM performance. Our analysis shows that classifier improvement primarily depends on the seed data selection, rather than iterative refinement using inferred samples. Interestingly, we find that intersecting high-quality data filtered by multiple classifiers consistently improves LLM performance.

2.5 FastText-based Quality Filtering

Current high-quality data classifiers are primarily divided into LLM-based (Penedo et al., 2024; Yu et al., 2025; Wang et al., 2024) and fastText-based (Li et al., 2024; Shao et al., 2024; Guo et al., 2024) methods. While LLM-based classifiers are effective, they need significantly higher inference costs. To address this, we adopt a fastText-based classifier, which significantly reduces inference costs while maintaining competitive performance under certain conditions. This approach not only minimizes resource consumption but also speeds up data filtering experiments. For instance, as shown in Table 2, processing 15T tokens with an LLM-based classifier requires approximately 6,000 H100 GPU hours, while fastText can complete the same task on a

²<https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu-llama3-annotations>

³<https://huggingface.co/datasets/BAAI/CCI3-HQ-Annotation-Benchmark>

non-GPU machine with just 80 CPUs in 1,000 hours, significantly improving efficiency. Notably, most of our large-scale experiments are conducted in a distributed manner using a Spark⁴ cluster.

Table 2: Comparison of inference costs for different model-based classifiers on 15T tokens

	LLM-based Classifier	fastText-based Classifier
GPU Used	✓	✗
CPU Used	✓	✓
Processing Time (Hours)	6,000	1,000

For data preprocessing, we implement several key steps, including removing redundant empty lines and extra spaces, stripping diacritics, and converting all English text to lowercase. Additionally, we adopt the DeepSeek-V2 tokenizer (Liu et al., 2024), which outperforms traditional tokenization methods (such as space-based tokenization for English and Jieba⁵ for Chinese). Meanwhile, we preserve structural information such as \n, \t, and \r. To ensure dataset integrity and balance, the final training set comprised 600K samples, evenly split between positive and negative examples.

For training details, we trained a fastText classifier with a vector dimension of 256, a learning rate of 0.1, a maximum word n-gram length of 3, a minimum word occurrence threshold of 5, and a total of 3 training epochs. Additionally, during inference, we maintain the default threshold of 0.5 to simplify operations and ensure experimental consistency, avoiding the need for additional tuning steps.

3 Experiments

In this section, we first detail the experimental settings in Section 3.1, including the training configuration, data composition, and evaluation metrics. Then, in Section 3.2, we present the overall experimental results, highlighting the performance comparisons between individual datasets and mixed datasets. These results demonstrate that *Ultra-FineWeb*, obtained through our efficient data filtering pipeline, exhibits superior quality to other datasets derived from the same data source, with the corresponding trained models achieving enhanced performance. Finally, in Section 3.3, we perform extensive ablation studies to further evaluate the effectiveness of *Ultra-FineWeb*, examining the impact of seed and recipe selection strategies for classifier training and the quality of the intersection of positive samples filtered by multiple classifiers.

3.1 Experimental Setting

Model Training Configuration. In our experiments, all models are trained using the open-source Megatron-LM library (Shoeybi et al., 2019). We utilize the MiniCPM-1.2B model architecture with the MiniCPM3-4B tokenizer. Each experiment involves training on 100B tokens (though the actual number is 104B tokens, calculated as $4096 \times 1024 \times 26000 = 104B$ tokens; for simplicity, we refer to it as 100B), allowing for comprehensive data performance validation within computationally efficient parameters. Key training parameters include a sequence length of 4096, weight decay of 0.1, and a gradient clipping threshold of 1.0. We employ a global batch size of 1,024 across 26,000 training steps. The learning rate follows a cosine decay schedule, with a warm-up phase of 1,000 steps. The initial learning rate is set to 1e-5, the maximum learning rate to 1e-2, and the final learning rate to 1e-3. To enhance training stability, we use Maximal Update Parameterization (MuP) (Yang et al., 2022). Additionally, we save a checkpoint every 1,000 steps (approximately 4B tokens) for analysis during the training process. Detailed model configurations are provided in Table 3, where *Params.*, *Vocab.*, d_m , d_{ff} , d_h , n_{head} , n_{kv} , and n_{Layer} represent the total number of non-embedding parameters, vocabulary size, model hidden dimension, feedforward layer bottleneck dimension, attention head dimension, number of queries, number of key/values, and the number of layers, respectively.

Table 3: Model Configurations for the MiniCPM-1.2B model.

Name	<i>Params.</i>	<i>Vocab.</i>	d_m	d_{ff}	d_h	n_{head}	n_{kv}	n_{Layer}
MiniCPM-1.2B	1,247,442,432	73448	1,536	3,840	64	24	8	52

⁴<https://spark.apache.org/>

⁵<https://pypi.org/project/jieba/>

Dataset Composition. We conduct two types of experiments for evaluating the datasets generated by our pipeline:

- **Individual Data Experiments:** We perform isolated training runs using single datasets, facilitating direct comparisons between differently processed data from identical sources. For English datasets, FineWeb is chosen as the source dataset, and comparisons are made with FineWeb-edu and Ultra-FineWeb-en. For Chinese datasets, Chinese FineWeb is selected with comparisons to Chinese FineWeb-edu-v2 and Ultra-FineWeb-zh. In the ablation studies, we primarily use individual data experiments for analysis.
- **Mixed Data Experiments:** Similar to the CCI3-HQ (Wang et al., 2024) experiment, we use a mix of 60% English data, 30% Chinese data, and 10% code data. The English-Chinese comparisons involve three dataset combinations: (1) FineWeb and Chinese FineWeb, (2) FineWeb-edu and Chinese FineWeb-edu-v2, and (3) Ultra-FineWeb-en and Ultra-FineWeb-zh. The code data is sourced exclusively from the StarCoder-v2 dataset (Lozhkov et al., 2024), maintaining consistent proportions across all experimental conditions.

Evaluation Metrics. We employ the Lighteval (Fourrier et al., 2023) library for model evaluation, mirroring the setup used with FineWeb (Penedo et al., 2024) and CCI3-HQ (Wang et al., 2024). All evaluation metrics are based on a zero-shot setting. The evaluation metrics include:

- *Average_{English}*: Average score across standard English metrics including MMLU (Hendrycks et al., 2020), ARC-C (Clark et al., 2018), ARC-E (Clark et al., 2018), CommonSenseQA (Talmor et al., 2018), HellaSwag (Zellers et al., 2019), OpenbookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), and Winogrande (Sakaguchi et al., 2021).
- *Average_{Chinese}*: Average score of Chinese metrics, including C-Eval (Huang et al., 2023) and CMMLU (Li et al., 2023).
- *Average*: The combined average score of all the above evaluation metrics.

3.2 Overall Results

Individual Dataset Results. We compare the performance of models trained on 100B tokens using data extracted from the FineWeb and Chinese FineWeb sources, using three different approaches: raw data, LLM-based classifiers (-edu), and the fastText-based classifier trained via the Efficient Data Filtering Pipeline (Ultra-). As shown in Tables 4 and 5, on the English Metrics, Ultra-FineWeb-en demonstrates significant improvements in performance on multiple tasks, including MMLU, ARC-C, ARC-E, CommonSenseQA, and OpenBookQA. Specifically, Ultra-FineWeb outperforms FineWeb in these tasks, with only a slight drop of 0.15 percentage points (*pp*) in HellaSwag compared to FineWeb, but a 0.6*pp* improvement over FineWeb-edu. The English average score for Ultra-FineWeb-en (45.891*pp*) is 3.61*pp* higher than that of FineWeb (42.287*pp*) and 1.3*pp* higher than FineWeb-edu (44.560*pp*). On the Chinese metrics, Ultra-FineWeb-zh also outperforms both FineWeb-zh and FineWeb-edu-zh on C-Eval and CMMLU. Specifically, Ultra-FineWeb-zh improves by 0.31*pp* and 3.65*pp* over Chinese FineWeb and Chinese FineWeb-edu-v2 on C-Eval and CMMLU, respectively, and by 0.09*pp* and 0.13*pp* compared to FineWeb-edu-zh. The Chinese average score for Ultra-FineWeb-zh increases by 1.98*pp* and 0.61*pp*, respectively, compared to FineWeb-zh and FineWeb-edu-zh. These results indicate that our proposed High-Quality Data Filtering Pipeline significantly improves data quality, leading to notable improvements in model performance. Additionally, we evaluate the performance at each training checkpoint. As shown in Figure 2, Ultra-FineWeb-en surpasses both FineWeb and FineWeb-edu early in the training process, while Ultra-FineWeb-zh demonstrates a marked improvement in Chinese average scores after 40B tokens of training.

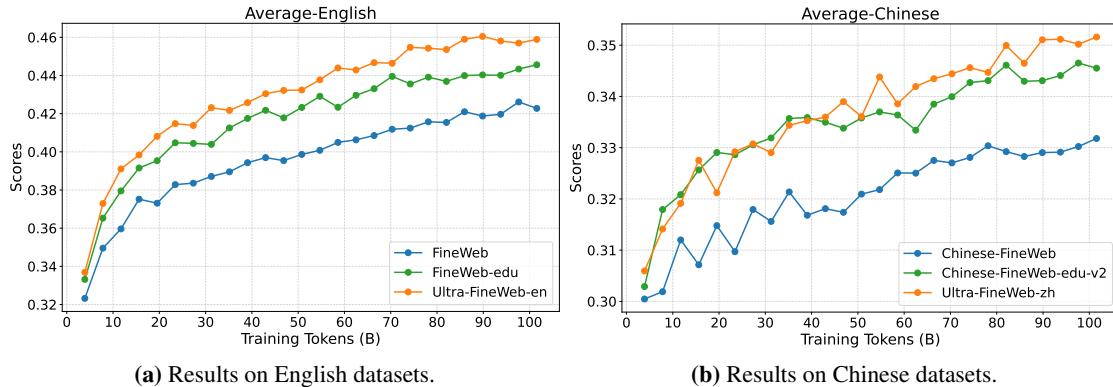
Mixed Dataset Results. In the mixed data experiments, we compare the model performance on different evaluation sets after training 100B tokens with the original data, LLM-based classifier-extracted edu data, and our Ultra-FineWeb dataset, using the same training configuration. As shown in Table 6, Ultra-FineWeb demonstrates significant performance improvements on multiple benchmarks. The average English score is 2.905*pp* higher than FineWeb_{mix} (41.366*pp*) and 0.538*pp* higher than FineWeb-edu_{mix} (43.733*pp*). For the Chinese evaluation set, Ultra-FineWeb achieves a 1.715*pp* advantage over than FineWeb_{mix} (32.01*pp*), while showing a marginal 0.025*pp* decrease compared to FineWeb-edu_{mix} (33.75*pp*). This minor discrepancy may stem from dataset weight setting or inherent training instability, warranting further investigation in future studies. The comprehensive analysis reveals Ultra-FineWeb’s superior performance over both baseline and LLM-filtered datasets, demonstrating significant overall score improvements. Despite task-specific fluctuations, Ultra-FineWeb, generated through our Efficient Data Filtering Pipeline, consistently delivers effective performance enhancements. The line charts of checkpoint evaluations are shown in Figure 3.

Table 4: Comparison of individual results on English datasets.

Metrics	FineWeb	FineWeb-edu	Ultra-FineWeb-en
MMLU	28.84	31.80 _{+2.96}	32.24_{+3.4}
ARC-C	25.17	34.56 _{+9.39}	35.67_{+10.5}
ARC-E	59.18	69.95 _{+10.77}	70.62_{+11.44}
CommonSenseQA	34.32	31.53 _{-2.79}	36.45_{+2.13}
HellaSwag	42.91	42.17 _{-0.74}	42.76 _{-0.15}
OpenbookQA	22.20	25.20 _{+3.00}	26.20_{+4.00}
PIQA	73.29	72.14 _{-1.15}	73.67_{+0.38}
SIQA	38.95	38.13 _{-0.82}	39.61_{+0.66}
Winogrande	55.64	55.56 _{-0.08}	55.80_{+0.16}
<i>Average_{English}</i>	42.278	44.560 _{+2.282}	45.891_{+3.613}

Table 5: Comparison of individual results on Chinese datasets.

Metrics	Chinese-FineWeb	Chinese-FineWeb-edu-v2	Ultra-FineWeb-zh
C-Eval	33.95	34.17 _{+0.22}	34.26_{+0.31}
CMMLU	32.41	34.93 _{+2.52}	36.06_{+3.65}
<i>Average_{Chinese}</i>	33.18	34.55 _{+1.370}	35.16_{+1.980}



(a) Results on English datasets.

(b) Results on Chinese datasets.

Figure 2: Average scores at each checkpoint for different individual datasets.

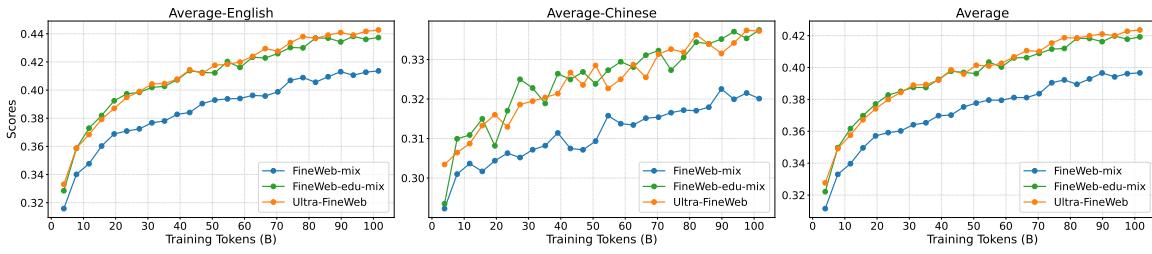
In the early training phases, Ultra-FineWeb and FineWeb-edu_{mix} exhibit comparable performance, but both outperform FineWeb_{mix}. Notably, Ultra-FineWeb starts to surpass FineWeb-edu_{mix} after training approximately 60B tokens. As for Chinese evaluation metrics, both Ultra-FineWeb and FineWeb-edu_{mix} demonstrate training fluctuations while maintaining substantial advantages over FineWeb_{mix} throughout the training process.

3.3 Ablation Study and Analysis

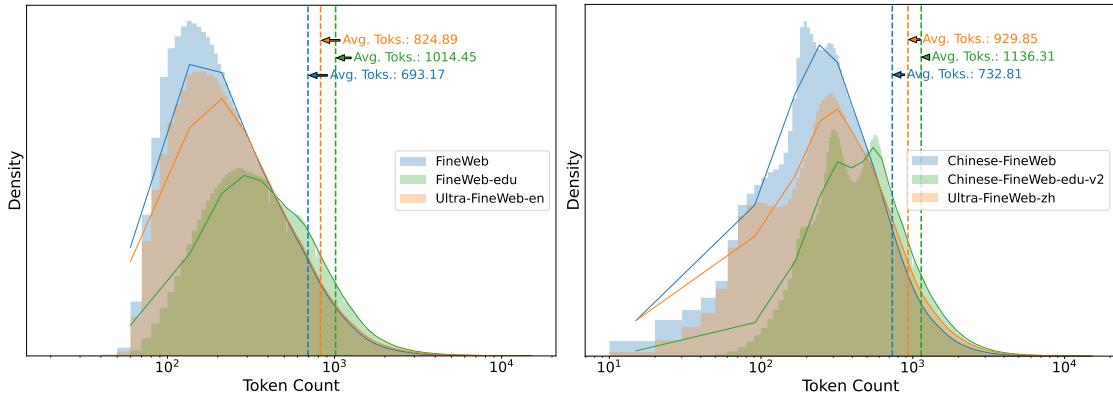
Analysis of Token Length Distributions. We first analyze the token length distributions across different datasets, as shown in Figure 4. For the English datasets, the token length distributions of Ultra-FineWeb-en and FineWeb are quite similar, while FineWeb-edu exhibits a rightward shift, indicating that the classifier tends to extract longer tokens. In terms of average token length, FineWeb has the shortest average, followed by Ultra-FineWeb, with FineWeb-edu having the longest average token length. For the Chinese datasets, Ultra-FineWeb-zh and Chinese FineWeb exhibit similar token length distributions, while Chinese FineWeb-edu-v2 also shows a rightward shift. The average token length follows the order: Chinese FineWeb < Chinese FineWeb-edu-v2 < Ultra-FineWeb-zh. We believe these differences may stem from the inherent preference of LLM-based models, which tend to favor longer tokens in their scoring. Additionally, this phenomenon might be further influenced by training recipes, as LLM-based models label data from the same source, typically

Table 6: Comparison of results on mixed datasets.

Metrics	FineWeb _{mix}	FineWeb-edu _{mix}	Ultra-FineWeb
MMLU	28.50	30.95 _{+2.45}	30.94 _{+2.44}
ARC-C	24.15	32.34 _{+8.19}	33.36 _{+9.21}
ARC-E	55.60	67.13 _{+11.53}	67.97 _{+12.37}
CommonSenseQA	36.20	35.79 _{-0.41}	37.18 _{+0.98}
HellaSwag	40.28	40.21 _{-0.07}	39.65 _{-0.63}
OpenbookQA	21.60	23.80 _{+2.20}	24.40 _{+2.80}
PIQA	71.11	71.22 _{+0.11}	70.08 _{-1.03}
SIQA	39.76	39.20 _{-0.56}	40.48 _{+0.72}
Winogrande	55.09	52.96 _{-2.13}	54.38 _{-0.71}
C-Eval	33.79	34.32 _{+0.53}	34.10 _{+0.31}
CMMLU	30.23	33.18 _{+2.95}	33.35 _{+3.12}
Average _{English}	41.366	43.733 _{+2.367}	44.271 _{+2.905}
Average _{Chinese}	32.010	33.750 _{+1.740}	33.725 _{+1.715}
Average	39.665	41.918 _{+2.253}	42.354 _{+2.689}

**Figure 3:** Average scores at each checkpoint for different mixed datasets.

assigning lower scores to shorter texts and higher scores to longer ones, leading classifiers to favor longer texts. In contrast, our data seeds are more diverse, making the classifiers less focused on token length, which results in the token length distribution of our extracted data aligning more closely with the original source data.



(a) Comparison of token length distributions on English datasets.
(b) Comparison of token length distributions on Chinese datasets.

Figure 4: Comparison of token length distributions across different datasets.

Loss and Performance Estimation Results. We use the performance estimation methods proposed in Xiao et al. (2024) for further analysis and verification of the effectiveness of Ultra-FineWeb. First, we establish the standard configuration in Xiao et al. (2024) as the baseline. Specifically, we adopt the MiniCPM-3-4B (Hu et al., 2024) training corpus, applying models across six scales (0.005B, 0.03B, 0.1B, 0.2B, 0.4B, 0.8B), and

train with six token configurations ($10, 15, 20, 30, 40, 60 \times N$, where N represents the model parameter size). Based on these 36 models, we compute and plot the compute ($= 6ND$)-Loss curve, and subsequently predict the performance of each model using the Loss-Performance curve from the Densing Law. This analysis is performed on MMLU (Hendrycks et al., 2020), BBH (Suzgun et al., 2022), MATH (Hendrycks et al., 2021), MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021), C-Eval (Huang et al., 2023), and CMMLU (Li et al., 2023) evaluation metrics. Next, we replace the “High-Quality” data in the baseline with ***Ultra-FineWeb*** and repeat the experiment, performing Loss Estimation. Finally, through this two-step estimation, we predict the performance of an 8B model trained on 8T tokens. The loss values and estimated results are shown in Table 7, with the Loss-Performance curve shown in Figure 5. Experimental results demonstrate that using Ultra-FineWeb significantly reduces the loss for metrics such as MMLU, MATH, C-Eval, and CMMLU, thereby improves model performance.

Table 7: Loss values and estimated performance for 8B model trained on 8T tokens.

Metrics	Baseline		Ultra-FineWeb	
	Loss	Estimate Acc.	Loss	Estimate Acc.
MMLU	0.182	70.84	0.143 _{-0.039}	85.60 _{+14.76}
BBH	0.097	56.70	0.092 _{-0.005}	60.48 _{+3.78}
MATH	0.225	25.96	0.162 _{-0.063}	59.05 _{+33.09}
MBPP	0.175	84.91	0.176 _{+0.001}	84.87 _{-0.04}
HumanEval	0.119	48.18	0.113 _{-0.006}	54.81 _{+6.63}
C-Eval	0.244	60.44	0.226 _{-0.018}	69.33 _{+8.89}
CMMLU	0.243	66.02	0.226 _{-0.017}	73.75 _{+7.73}
<i>Average</i>	0.189	42.40	0.174 _{-0.015}	49.85 _{+7.45}

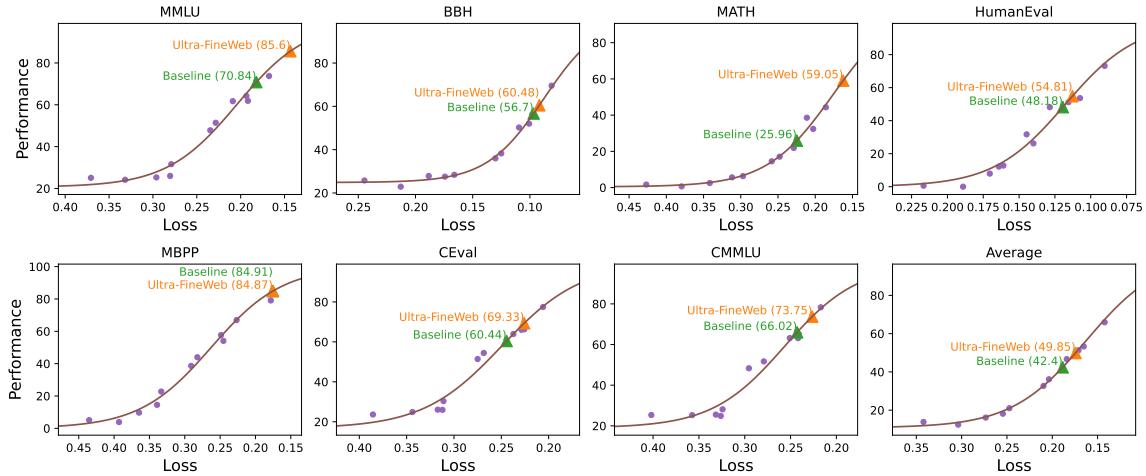


Figure 5: Loss-performance curve: Showing the estimated performance of an 8B model trained on 8T tokens using baseline and replacing high-quality data with Ultra-FineWeb.

Ablation Study on Multi-Source Seed Selection. To verify the impact of selecting multi-source seed on the robustness of the classifier during the efficient data filtering pipeline process, we choose DCLM-Pool (Li et al., 2024) as the English data source and MAP-CC (Du et al., 2024) as the Chinese data source for verification. In the experiment, we compare the performance of models trained with original data, LLM-based classifier (-edu), and data extracted by our classifier (Ultra-) on different evaluation sets. Notably, due to the unavailability of an open-source LLM-based classifier for Chinese-FineWeb-edu, we only compare the performance difference between the original MAP-CC data and the data extracted by our classifier (Ultra-MAP-CC). As detailed in Tables 8 and 9, Ultra-DCLM demonstrates superior performance over both DCLM-Pool and DCLM-edu across multiple English evaluation tasks. The English average score for Ultra-DCLM (47.252pp) shows a 1.671pp improvement over DCLM-Pool (45.581pp) and a 0.658pp advantage over DCLM-edu (46.594pp),

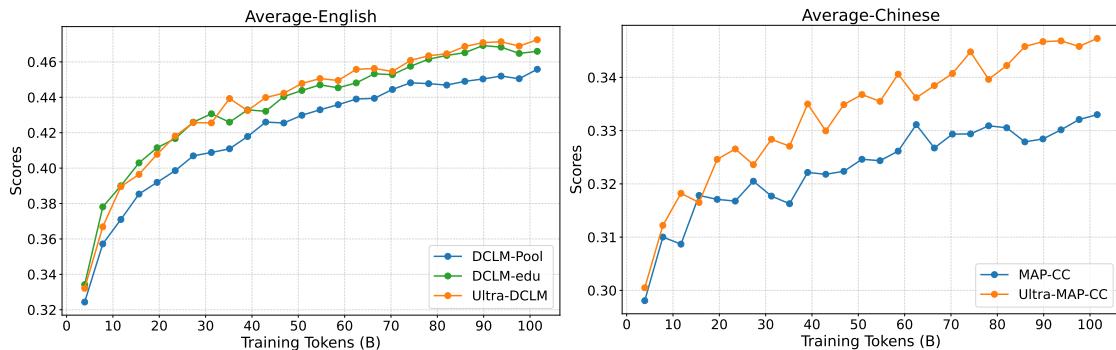
with particularly notable gains in MMLU, ARC-C, and OpenbookQA metrics. For Chinese evaluations, Ultra-MAP-CC also exhibits significant enhancements, especially in CMMLU with a $2.8pp$ increase, achieving an overall $1.43pp$ improvement over the original dataset. These results demonstrate that our classifier remains highly robust and effective even in non-homogeneous data scenarios, further confirming the positive impact of the multi-source seed selection strategy on improving classifier robustness and performance. Figure 6 presents the evaluation results at each checkpoint during training. In the early stages of training, the performance of Ultra-DCLM and DCLM-edu is similar, but both outperform DCLM-Pool significantly. When training reaches 30B tokens, Ultra-DCLM begins to surpass DCLM-edu. For the Chinese evaluation sets, Ultra-MAP-CC significantly outperforms MAP-CC from the early stages of training.

Table 8: Comparison of results on DCLM-Pool-based datasets.

Metrics	DCLM-Pool	DCLM-edu	Ultra-DCLM
MMLU	31.45	$34.07_{-2.62}^{+2.62}$	$34.33_{-2.88}^{+2.88}$
ARC-C	31.48	$37.71_{-6.23}^{+6.23}$	$38.48_{-7.00}^{+7.00}$
ARC-E	66.08	$73.40_{-7.32}^{+7.32}$	$72.77_{-6.69}^{+6.69}$
CommonSenseQA	41.52	$39.72_{-1.80}^{+1.80}$	$40.70_{-0.82}^{+0.82}$
HellaSwag	44.28	$41.77_{-2.51}^{+2.51}$	$43.31_{-0.97}^{+0.97}$
OpenbookQA	25.00	$26.60_{-1.60}^{+1.60}$	$27.40_{-2.40}^{+2.40}$
PIQA	73.67	$70.73_{-2.94}^{+2.94}$	$73.89_{-0.22}^{+0.22}$
SIQA	40.79	$39.00_{-1.79}^{+1.79}$	$39.51_{-1.28}^{+1.28}$
Winogrande	55.96	$56.35_{-0.39}^{+0.39}$	$54.88_{-1.08}^{+1.08}$
<i>Average_{English}</i>	45.581	$46.594_{-1.013}^{+1.013}$	$47.252_{-1.671}^{+1.671}$

Table 9: Comparison of results on MAP-CC-based datasets.

Metrics	MAP-CC	Ultra-MAP-CC
C-Eval	34.58	$34.64_{-0.06}^{+0.06}$
CMMLU	32.02	$34.82_{-2.80}^{+2.80}$
<i>Average_{Chinese}</i>	33.300	$34.730_{-1.430}^{+1.430}$



(a) Comparison of results on DCLM-Pool-based datasets at each checkpoint. (b) Comparison of results on MAP-CC-based datasets at each checkpoint.

Figure 6: Average scores at each checkpoint during training for different source data.

Ablation Study of Multi-Turn Training Recipes. To verify the impact of multiple iterations on classifier performance, we implement three rounds of iterations for both English and Chinese classifiers. The initial iteration utilizes the selected high-quality seed data for positive samples and multi-source original data for negative samples. The second iteration involves using the classifier from the first round to process the negative samples, and the inferred positive and negative samples are incorporated into the next round of training data. The third iteration involves updating the classifier with more precisely identified samples from the second

round. Experimental results (Tables 10 and 11) indicate that second-iteration classifiers achieved superior performance across multiple tasks compared to the first-iteration. Notably, English classifiers demonstrate significant improvements in MMLU, ARC-C, and OpenbookQA tasks, with an average score increase of 3.613 percentage points (*pp*) over both the first iteration and original FineWeb dataset, reaching 45.89*pp*. However, the third iteration, which focused solely on updating samples from original source data, failed to yield additional performance gains. In fact, there were slight declines in some tasks, such as HellaSwag and PIQA. For the Chinese data, the second iteration of Ultra-FineWeb-zh also shows notable improvements in CMMLU and C-Eval. However, similar to the English results, the third iteration provided only marginal overall improvements, with no significant gains in specific tasks. This suggests that iterative sample refinement through enhanced classifiers alone is insufficient for achieving further performance improvements.

Table 10: Comparison of results on English datasets with multiple iterations.

Metrics	FineWeb	fastText-en-v1	Ultra-FineWeb-en	fastText-en-v3
MMLU	28.84	32.30 _{+3.46}	32.24 _{+3.40}	32.29 _{+3.45}
ARC-C	25.17	35.67 _{+10.50}	35.67 _{+10.50}	35.07 _{+9.9}
ARC-E	59.18	70.33 _{+11.15}	70.62 _{+11.44}	70.54 _{+11.36}
CommonSenseQA	34.32	32.27 _{-2.05}	36.45 _{+2.13}	36.55 _{+2.23}
HellaSwag	42.91	42.82 _{-0.09}	42.76 _{-0.15}	42.62 _{-0.29}
OpenbookQA	22.20	24.40 _{+2.20}	26.20 _{+4.00}	26.20 _{+4.00}
PIQA	73.29	72.09 _{-1.20}	73.67 _{+0.38}	72.53 _{-0.76}
SIQA	38.95	38.59 _{-0.36}	39.61 _{+0.66}	39.41 _{+0.46}
Winogrande	55.64	55.09 _{-0.55}	55.80 _{+0.16}	55.92 _{+0.28}
<i>Average_{English}</i>	42.278	44.840 _{+2.562}	45.891 _{+3.613}	45.681 _{+3.403}

Table 11: Comparison of results on Chinese datasets with multiple iterations.

Metrics	Chinese-FineWeb	fastText-zh-v1	Ultra-FineWeb-zh	fastText-zh-v3
C-Eval	33.95	33.63 _{-0.32}	34.26 _{+0.31}	34.26 _{+0.31}
CMMLU	32.41	35.82 _{+3.41}	36.06 _{+3.65}	35.07 _{+2.66}
<i>Average_{Chinese}</i>	34.035	35.390 _{+1.355}	35.875 _{+1.840}	34.26 _{+0.225}

Ablation Study on Classifier Inference Intersection. To investigate the impact of intersecting positive samples from multiple classifiers on LLM performance, we conduct experiments using the intersection of classifier-inferred positive samples for model training. As demonstrated in Tables 12 and 13, the model trained on Ultra-FineWeb-en_{inter} exhibits substantial performance gains across multiple English metrics. Compared to the Ultra-FineWeb-en, the score improved by 0.447*pp*, with the most significant improvements observed in tasks such as MMLU, ARC-C, ARC-E, and OpenbookQA. Similarly, for Chinese metrics, the model trained on Ultra-FineWeb-zh_{inter} also showed notable performance gains, with the overall Chinese average score increasing from 35.16*pp* to 36.455*pp*. In particular, the score in CMMLU improved by 1.8*pp* compared to Ultra-FineWeb-zh. Additionally, we visualize the evaluation scores at each checkpoint during training, as shown in Figure 7, where the model using intersection data consistently maintain the highest score throughout training. These results indicate that combining the inference results from multiple classifiers, particularly through intersecting positive sample data, can significantly further enhance model performance, yielding significant improvements across key metrics.

4 Related Work

The success of LLMs largely depends on the availability of large-scale, high-quality pretraining corpora, which provide models with rich knowledge and reasoning capabilities. Common Crawl (Crawl, 2007) has served as the foundation data source for LLM development, and to meet the growing data requirements for training larger models, a vast amount of pretraining corpora have been made open source. Early efforts, such as C4 (Raffel et al., 2020) with 160B tokens and Pile (Gao et al., 2020) with 300B tokens, provide critical resources for early model pretraining. In recent years, substantially larger corpora have emerged, including RefinedWeb (Penedo et al., 2023) with 600B tokens, Dolma (Soldaini et al., 2024) with 3T tokens,

Table 12: Comparison of results on English datasets using the intersection of positive samples inferred by multiple classifiers.

Metrics	FineWeb	FineWeb-edu	Ultra-FineWeb-en	Ultra-FineWeb-en _{inter}
MMLU	28.84	31.80 _{+2.96}	32.24 _{+3.40}	33.37 _{+4.53}
ARC-C	25.17	34.56 _{+9.39}	35.67 _{+10.50}	38.31 _{+13.14}
ARC-E	59.18	69.95 _{+10.77}	70.62 _{+11.44}	73.48 _{+14.30}
CommonSenseQA	34.32	31.53 _{-2.79}	36.45 _{+2.13}	36.94 _{+2.62}
HellaSwag	42.91	42.17 _{-0.74}	42.76 _{-0.15}	41.39 _{-1.52}
OpenbookQA	22.20	25.20 _{+3.00}	26.20 _{+4.00}	28.60 _{+6.40}
PIQA	73.29	72.14 _{-1.15}	73.67 _{+0.38}	71.16 _{-2.13}
SIQA	38.95	38.13 _{-0.82}	39.61 _{+0.66}	39.41 _{+0.46}
Winogrande	55.64	55.56 _{-0.08}	55.80 _{+0.16}	54.38 _{-1.26}
<i>Average_{English}</i>	42.278	44.560 _{+2.282}	45.891 _{+3.613}	46.338 _{+4.06}

Table 13: Comparison of results on Chinese Datasets using the intersection of positive samples inferred by multiple classifiers.

Metrics	Chinese-FineWeb	Chinese-FineWeb-edu-v2	Ultra-FineWeb-zh	Ultra-FineWeb-zh _{inter}
C-Eval	33.95	34.17 _{+0.22}	34.26 _{+0.31}	35.05 _{+1.1}
CMMLU	32.41	34.93 _{+2.52}	36.06 _{+3.65}	37.86 _{+5.45}
<i>Average_{Chinese}</i>	33.180	34.550 _{+1.37}	35.160 _{+1.98}	36.455 _{+3.275}

FineWeb (Penedo et al., 2024) with 15T tokens, RedPajama-v2 (Weber et al., 2025) with 30T tokens, and DCLM (Li et al., 2024) with 240T tokens, significantly advancing LLM development, fostering community collaboration, and establishing new benchmarks for innovation. Meanwhile, Chinese pretraining corpora have also been rapidly developed, such as ChineseWebText (Chen et al., 2023) with 50B tokens, WuDao (BAAI, 2023) with 120B tokens, IndustryCorpus2 (Shi et al., 2024) with 200B tokens, and CCI3 (Wang et al., 2024) with 200B tokens. However, despite progress in traditional data processing methods (such as heuristic filtering and deduplication) during the early stages, the processed data still often contains noise and unstructured content. With the continuous scaling up of models and increasing demands for data quality, these methods have become insufficient to meet current requirements.

To address these challenges, model-driven data filtering strategies have gradually become an effective approach to improving data quality in recent years. These approaches are primarily implemented during the final stages of large-scale data preprocessing, aiming to filter high-quality and high-value samples from massive datasets to further enhance model performance. Traditional quality filtering techniques (Penedo et al., 2024; Wang

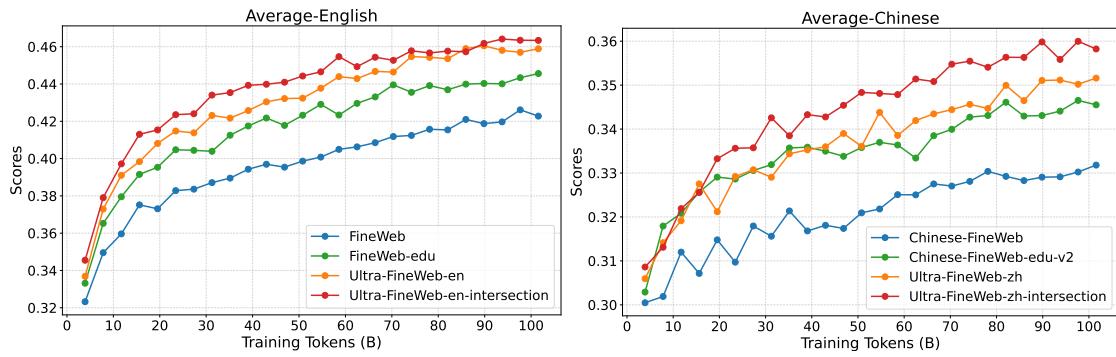


Figure 7: Average scores at each checkpoint during training for different datasets: Compared with using the intersection of positive samples inferred by multiple classifiers.

et al., 2024; Li et al., 2024; Yu et al., 2025) typically train classifiers to distinguish between high-quality data (such as textbook text) and low-quality data (such as raw web text), subsequently filtering out samples with lower inference scores. Additionally, data filtering methods based on perplexity (Muennighoff et al., 2024; Wenzek et al., 2019), and strategies using pre-trained LLMs to evaluate multiple dimensions of data quality through prompts (Sachdeva et al., 2024; Wettig et al., 2024), have been introduced. These advancements have greatly expanded the range of data filtering methods available.

The common trend of these methods is to obtain higher-quality data by reducing computational costs. By optimizing the filtering process and reducing inference resource consumption, not only is dataset quality improved, but data processing efficiency is also accelerated. This optimization enables LLMs to access superior training corpora, facilitating enhanced model performance with reduced training token requirements.

5 Conclusion

In this paper, we construct a higher-quality ***Ultra-FineWeb*** dataset (including English data *Ultra-FineWeb-en*, approximately 1T tokens, and Chinese data *Ultra-FineWeb-zh*, approximately 120B tokens, totaling approximately 1.1T tokens). This dataset is based on the FineWeb and Chinese FineWeb datasets, utilizing our proposed efficient data filtering pipeline. Through rigorous experimental evaluations, we demonstrate that Ultra-FineWeb-en and Ultra-FineWeb-zh outperform FineWeb-edu and Chinese FineWeb-edu-v2 when used for small-scale model training from scratch. Additionally, we show the effectiveness of the high-quality data filtered by our classifier on the DCLM-Pool and MAP-CC datasets, further confirming the reliability and effectiveness of our proposed pipeline. These results indicate that classifiers based on our efficient data filtering pipeline can select higher-quality data with reduced computational cost, thereby improving model training performance. We provide a detailed description of the implementation of our efficient data filtering pipeline, especially the efficient verification strategy driven by classifiers in the pipeline. This strategy enables reliable assessment of training data impact on LLM performance while maintaining minimal computational requirements. Furthermore, we present detailed methodologies for classifier seed data selection, training recipes, and FastText model training configuration, ensuring experimental reproducibility and result transparency. This study aims to provide novel insights and methodologies for high-quality data filtering, offering valuable references for data quality optimization in future LLM training processes, and contributing to the further development of LLMs.

6 Limitations and Future Directions

Some key limitations of our work are as follows. Due to time and resource constraints, we did not conduct an comprehensive analysis of classifier inference thresholds. Although the default threshold ($thr = 0.5$) is effective in filtering higher-quality data, in future work, we plan to explore the impact of different threshold ranges on data quality to further optimize the filtering strategy. Specifically, during multiple iterations, we can experiment with different thresholds to refine the filtering process, enabling more precise selection of data at different quality levels for the next round of classifier training. This approach will help dynamically adjust the filtering strategy and progressively optimize the dataset, thereby improving classifier performance and data quality. Additionally, this paper primarily focuses on general high-quality datasets and validates the effectiveness of the efficient data filtering pipeline. In the future, we hope to extend this method to more specialized domains, such as mathematics, code, law, and other technical fields, to meet the requirements of different scenarios. This would contribute to the creation of more targeted high-quality datasets, enhancing model performance on specific tasks. Furthermore, current work typically evaluates data quality through model training results, which is heavily reliant on the performance of the trained models and lacks more objective and systematic quality metrics. Therefore, in future research, we aim to develop some quantifiable data quality evaluation standards or tools to provide multidimensional measurements of data quality. This would further enhance the precision and operability of data filtering, providing a more scientific basis for building high-quality datasets. These efforts will further improve our data filtering methods and provide higher-quality data support for the training of large-scale language models.

References

- Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.
- BAAI. Wudao corpus, 2023. URL <https://data.baai.ac.cn/details/WuDaoCorporaText>.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.
- Zheng Cai, Maosong Cao, Haojong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.
- Jianghao Chen, Pu Jian, Tengxiao Xi, Dongyi Yi, Qianlong Du, Chenglin Ding, Guibo Zhu, Chengqing Zong, Jinqiao Wang, and Jiajun Zhang. ChineseWebText: Large-scale high-quality chinese web text extracted with effective evaluation model. *arXiv preprint arXiv:2311.01149*, 2023.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harry Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. <https://github.com/open-compass/opencompass>, 2023.
- Common Crawl. Common crawl. <https://commoncrawl.org>, 2007.
- Xinrun Du, Zhouliang Yu, Songyang Gao, Ding Pan, Yuyang Cheng, Ziyang Ma, Ruibin Yuan, Xingwei Qu, Jiaheng Liu, Tianyu Zheng, et al. Chinese tiny llm: Pretraining a chinese-centric large language model. *arXiv preprint arXiv:2404.04167*, 2024.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Clémentine Fourrier, Nathan Habib, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023. URL <https://github.com/huggingface/lighteval>.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The Pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming—the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.
- Daya Guo, Dejian Yang, Huawei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, et al. Pre-trained models: Past, present and future. *AI Open*, 2:225–250, 2021.

- Zhongjiang He, Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. Telechat technical report. *arXiv preprint arXiv:2401.03804*, 2024.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*, 2023.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou, and Tomas Mikolov. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*, 2016.
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2023.
- Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *arXiv preprint arXiv:2406.11794*, 2024.
- Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.
- Yidong Liu, FuKai Shang, Fang Wang, Rui Xu, Jun Wang, Wei Li, Yao Li, and Conghui He. Michao-huafen 1.0: A specialized pre-trained corpus dataset for domain-specific large models. *arXiv preprint arXiv:2309.13079*, 2023.
- Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, et al. Starcoder 2 and the stack v2: The next generation. *arXiv preprint arXiv:2402.19173*, 2024.
- Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. Llm4sr: A survey on large language models for scientific research. *arXiv preprint arXiv:2501.04306*, 2025.
- Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. Exploring the limit of outcome reward for learning mathematical reasoning. *arXiv preprint arXiv:2502.06781*, 2025.
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.
- Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.
- Jiantao Qiu, Haijun Lv, Zhenjiang Jin, Rui Wang, Wenchang Ning, Jia Yu, ChaoBin Zhang, Zhenxiang Li, Pei Chu, Yuan Qu, et al. Wanjuan-cc: A safe and high-quality open-sourced english webtext dataset. *arXiv preprint arXiv:2402.19282*, 2024.
- Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*, 2024.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialqa: Commonsense reasoning about social interactions, 2019.
- Zhihong Shao, Peiyi Wang, Qihaoo Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- Xiaofeng Shi, Lulu Zhao, Hua Zhou, and Donglin Hao. Industrycorpus2, 2024. URL <https://huggingface.co/datasets/BAAI/IndustryCorpus2>.
- Mohammad Shoeybi, Mostafa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.
- Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Arthur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.
- Zifan Song, Yudong Wang, Wenwei Zhang, Kuikun Liu, Chengqi Lyu, Demin Song, Qipeng Guo, Hang Yan, Dahua Lin, Kai Chen, et al. Alchemicoder: Harmonizing and eliciting code capability by hindsight tuning on multi-source data. *arXiv preprint arXiv:2405.19265*, 2024.
- Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- InternLM Team. InternLM: A multilingual language model with progressively enhanced capabilities, 2023.
- Teknium. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants, 2023. URL <https://huggingface.co/datasets/teknium/OpenHermes-2.5>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- Liangdong Wang, Bo-Wen Zhang, Chengwei Wu, Hanyu Zhao, Xiaofeng Shi, Shuhao Gu, Jijie Li, Quanyue Ma, TengFei Pan, and Guang Liu. Cci3. 0-hq: a large-scale chinese dataset of high quality designed for pre-training large language models. *arXiv preprint arXiv:2410.18505*, 2024.

Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. Redpajama: an open dataset for training large language models. *Advances in Neural Information Processing Systems*, 37:116462–116492, 2025.

Tianwen Wei, Liang Zhao, Lichang Zhang, Bo Zhu, Lijie Wang, Haihua Yang, Biye Li, Cheng Cheng, Weiwei Lü, Rui Hu, et al. Skywork: A more open bilingual foundation model. *arXiv preprint arXiv:2310.19341*, 2023.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

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

Chaojun Xiao, Jie Cai, Weilin Zhao, Guoyang Zeng, Xu Han, Zhiyuan Liu, and Maosong Sun. Densing law of llms. *arXiv preprint arXiv:2412.04315*, 2024.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Yijiong Yu, Ziyun Dai, Zekun Wang, Wei Wang, Ran Chen, and Ji Pei. Opencsg chinese corpus: A series of high-quality chinese datasets for llm training. *arXiv preprint arXiv:2501.08197*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Kaiyan Zhang, Sihang Zeng, Ermo Hua, Ning Ding, Zhang-Ren Chen, Zhiyuan Ma, Haoxin Li, Ganqu Cui, Biqing Qi, Xuekai Zhu, et al. Ultramedical: Building specialized generalists in biomedicine. *arXiv preprint arXiv:2406.03949*, 2024.

A Implementation Details for Efficient Verification

All Efficient Verification experiments are trained using the open-source Megatron-LM library. We utilize the MiniCPM-1.2B model architecture (as shown in Table 1) with the MiniCPM-3-4B tokenizer. The model training utilizes the MiniCPM-3-4B training corpus, and the WSD scheduler. We train the model from scratch with 1.1T tokens (1T for the stable stage and 0.1T for the decaying stage). Based on this pretrained model, we further perform a two-stage annealing training with 10B tokens, allocating 30% of the weight to the verification data, while keeping the remaining 70% for the default mixed data ratio. Key training parameters include a sequence length of 4096, weight decay of 0.1, and a gradient clipping threshold of 1.0. We employed a global batch size of 512. For larger datasets, we train for a total of 5000 steps (approximately 10B tokens). For smaller datasets, we compute the total training steps based on the actual token size of the data and typically allowed the validation data to undergo 3-5 training epochs (n_{epoch}).

The training steps calculation formula is as follows:

$$Total\ Iter = \max \left(\frac{Total\ Token}{Global\ BS \times Seq\ Len}, 5000 \right)$$

Where *Total Token* is calculated as:

$$Total\ Token = \frac{Curr\ Data\ Token \times n_{epoch}}{0.3}$$

To reduce the cost of baseline experiments, we typically choose training steps of 100, 500, 1,000, 2,500, or 5,000, balancing experimental accuracy and computational resource consumption. This means that for datasets of different scales, we dynamically adjust the training steps based on the data tokens required for training. It is important to note that the training steps for the baseline experiments are also dynamically adjusted based on the corresponding dataset size. Additionally, we set the warmup fraction to 0.1, and the annealing phase used an exponential decay approach, with the maximum learning rate to 1e-3 and the minimum learning rate to 5e-5. To enhance training stability, we use Maximal Update Parameterization (MuP).

B Results and Analysis of Efficient Verification

To verify the effectiveness of the efficient verification strategy, we use the FineWeb and FineWeb-edu datasets, training both on the efficient verification and from-scratch 100B token strategies, and compare the results. For evaluation, we use OpenCompass (Contributors, 2023) for the model trained with the efficient verification strategy, and Lighteval (Fourier et al., 2023) for the model trained from scratch with 100B tokens. The experimental results are shown in Table 14.

Table 14: Comparison of efficient verification strategy and from-scratch 100B token strategies on FineWeb and FineWeb-edu.

Metrics	Efficient Verification			100B From Scratch		
	FineWeb	FineWeb-edu	Diff.	FineWeb	FineWeb-edu	Diff.
MMLU	45.84	47.35	+1.51	28.84	31.80	+2.96
HellaSwag	57.72	56.99	-0.73	42.91	42.17	-0.74
ARC-C	38.98	39.66	+0.68	25.17	34.56	+9.39
ARC-E	57.67	59.08	+1.41	59.18	69.95	+10.77
PIQA	74.48	72.91	-1.57	73.29	72.14	-1.15
SIQA	43.55	43.35	-0.20	38.95	38.13	-0.82
Winogrande	56.67	55.56	-1.11	55.64	55.56	-0.08
OpenbookQA	66.80	69.40	+2.60	22.20	25.20	+3.00
<i>Average</i>	55.21	55.54	+0.33	41.00	43.00	+2.00

We can observe that the efficient verification strategy exhibited consistent trends across multiple evaluation tasks when compared to the from-scratch 100B model. For example, in metrics like MMLU, ARC-E, ARC-C,

and OpenbookQA, FineWeb-edu consistently outperformed FineWeb under both training paradigms. Similarly, for metrics such as HellaSwag, PIQA, SIQA, and Winogrande, FineWeb-edu showed performance degradation compared to FineWeb, regardless of training strategy. Overall, the efficient verification strategy quickly revealed the impact of the validation data on various evaluation dimensions and provided accurate feedback. This strategy significantly reduces computational resource requirements, enabling more efficient data quality assessment and optimization, ultimately enhancing model training effectiveness.