UTF8gbsn

CCI4.0: 增强大型语言模型推理能力的双语预训练数据集

Guang Liu[*] **Liangdong Wang** **Jijie Li**
Yang Yu **Yao Xu** **Jiabei Chen** **Yu Bai** **Feng Liao** **Yonghua Lin**
Data Research Team
Beijing Academy of Artificial Intelligence

Abstract

我们介绍了 CCI4.0，这是一个大型双语预训练数据集，旨在提供卓越的数据质量和多样化的人类类推理轨迹。CCI4.0 大约占用 35 TB 的磁盘空间，包含两个子数据集：CCI4.0-M2-Base 和 CCI4.0-M2-CoT。CCI4.0-M2-Base 结合了一个 5.2 TB 的经过精心整理的中文网页语料库，Nemotron-CC 中一个 22.5 TB 的英文子集，以及来自数学、维基、arxiv 和代码的多样化来源。尽管这些数据大多来源于完善的数据集，但各个领域的质量标准是动态的，处理这些数据需要广泛的专家经验和大量的人力。因此，我们提出了一个新的方法论来评估数据质量，主要基于以下三个步骤：两阶段去重、多分类器质量评分和领域感知的流畅性过滤。我们提取了 4.5 亿条 CoT（思维链）模板，称为 CCI4.0-M2-CoT。与从大型模型中提取 CoT 的方式不同，我们建议的分阶段 CoT 提取展示了多样化的推理模式，并显著降低了幻觉的可能性。实证评估表明，使用 CCI4.0 预训练的 LLMs 受益于更清晰、更可靠的训练信号，在下游任务中表现出一致的改进，特别是在数学和代码反思任务中。我们的结果强调了严谨的数据整理和人类思维模板在提升 LLM 性能方面的关键作用，为自动化处理预训练语料库提供了一些启示。

1 引言

In recent years, large language models (LLMs) have achieved remarkable success across a broad spectrum of natural language processing tasks, including text generation, translation, and sentiment analysis. A critical factor underpinning these advances is the availability and quality of large-scale pre-training data [24, 33, 16]. Pre-training data not only shapes the linguistic capabilities of LLMs but also plays a central role in determining their generalization and reasoning abilities across a wide range of downstream applications.

Despite growing efforts to curate and release open-source datasets for language model training [31, 30, 17], there remains a significant gap in the availability of high-quality and diverse large-scale corpora. Most existing resources are limited in either linguistic diversity or domain coverage, constraining the models' ability to generalize beyond narrow contexts. Furthermore, although real-world data is often prioritized for its authenticity, synthetic data has emerged as an important complement, particularly in fostering reasoning skills. Nevertheless, high-quality synthetic pre-training datasets—especially those that explicitly incorporate structured reasoning—are still notably lacking.

To address the aforementioned gaps and promote the advancement of data-centric development in large language models, we introduce CCI4.0-M2-Base: a high-quality and diverse bilingual corpus in Chinese and English. In particular, CCI4.0-M2-Base includes a large-scale, bilingual pretraining dataset (35T tokens) combining a Chinese corpus and Nemotron-CC's English data, with a

[*]Corresponding Author, liuguang@baai.ac.cn

meticulously-designed pipeline to yield a high-quality dataset to enhance LLM general and reasoning capabilities. Moreover, to enhance the diversity of the corpus, we additionally incorporate high-quality source data from a wide range of domains, including web pages, code, mathematics, academic papers, and encyclopedias.

Furthermore, considering that the reasoning capabilities of LLMs are primarily developed during the pretraining phase [15, 39, 1], we provide CCI4.0-M2-CoT, which integrates 4.5 billion human thinking templates synthesized from high-quality samples using advanced techniques. While effective for general language understanding, traditional pretraining datasets often lack the specialized content needed to foster advanced reasoning skills, such as explicit representations of human thought processes or logical reasoning traces. To address this gap, we introduce CCI4.0-M2-CoT, whose templates are crafted to embed diverse reasoning patterns, strengthening the foundational reasoning abilities of LLMs, and decreasing the possibility of hallucination.

Experimental results validate the efficacy of CCI4.0, demonstrating substantial performance improvements on knowledge-based and reasoning-intensive benchmarks such as MMLU and ARC-Challenge, with notable gains in commonsense reasoning and mathematical problem-solving. These findings underscore the critical role of high-quality, diverse and reasoning-focused pretraining data in advancing LLMs' ability to tackle complex, multi-step reasoning tasks. By addressing the limitations of existing datasets, CCI4.0 sets a new benchmark for pretraining and paves the way for the development of more capable and versatile language models.

This paper makes the following core contributions:

- 介绍 CCI4.0-M2-Base：这是一个大规模的双语预训练数据集（3500 亿标记），结合了中文语料库、Nemotron-CC 的英语数据以及来源于不同领域的语料库，旨在增强大型语言模型的一般和推理能力。
- 合并 CCI4.0-M2-CoT：整合多样化的推理模板，包括 45 亿条合成的人类思维模板，嵌入多样化的推理模式以增强逻辑和常识推理并减少幻觉。
- 高级数据处理流程：包括去重、多分类器质量评分、流畅性过滤、CoT 合成以及隐私/有害内容处理的综合方法，确保高质量和多样化的数据整理。
- 实证验证：在如 MMLU 和 ARC-Challenge 的基准上显示了显著的性能提升，特别是在数学问题解决和常识推理方面，超越了如 Nemotron-CC-HQ 和 CCI3-HQ 这样的基线数据集。

Table 1: 数据集比较。列的缩写：Size（开源规模）、Multi-Src（多源）、Multi-Cls（多分类器）、Multi-Lang（多语言）、CoT-Syn（CoT 合成）。

Dataset	Size(TB)	Multi-Src	Multi-Cls	Multi-Lang	CoT-Syn
Pile	0.8	✓	×	✓	×
Dolma	11	✓	✓	×	×
RefindeWeb	0.6	×	×	✓	×
Redpajama (V2)	~ 120	×	✓	✓	×
FineWeb2	17.7	×	×	✓	×
Wanjuan1.0	0.19	×	✓	✓	×
FineWeb	44	×	×	×	×
Nemotron-CC	22	×	✓	×	×
CCI3.0	1	×	✓	×	×
CCI4.0	35	✓	✓	✓	✓

2 相关工作

To contextualize the development of CCI4.0, we compare it with existing large-scale pretraining datasets, as summarized in Table 1. The Pile (0.8T tokens) and Dolma (11T tokens) leverage multi-source data and multiple classifiers for quality assurance, supporting multilingual content but lacking Chain-of-Thought (CoT) synthesis, which limits their focus on reasoning enhancement. Similarly, datasets like Redpajama (V2) (120T tokens) and FineWeb2 (17.7T tokens) offer substantial scale and multilingual support, yet they do not incorporate CoT synthesis or multi-source diversity to the extent of CCI4.0. Nemotron-CC (10.4T tokens), a key component of CCI4.0' s English corpus, employs multiple classifiers but is monolingual and lacks CoT integration. In contrast, CCI4.0 (35T tokens) uniquely combines multi-source data, multilingual support (English and Chinese), multiple quality classifiers, and CoT synthesis, incorporating 4.5 billion human thinking templates to explicitly target reasoning capabilities. This comprehensive approach distinguishes CCI4.0 from prior datasets, addressing gaps in reasoning-focused pretraining data and setting a new benchmark for developing LLMs with enhanced logical and commonsense reasoning abilities.

3 方法

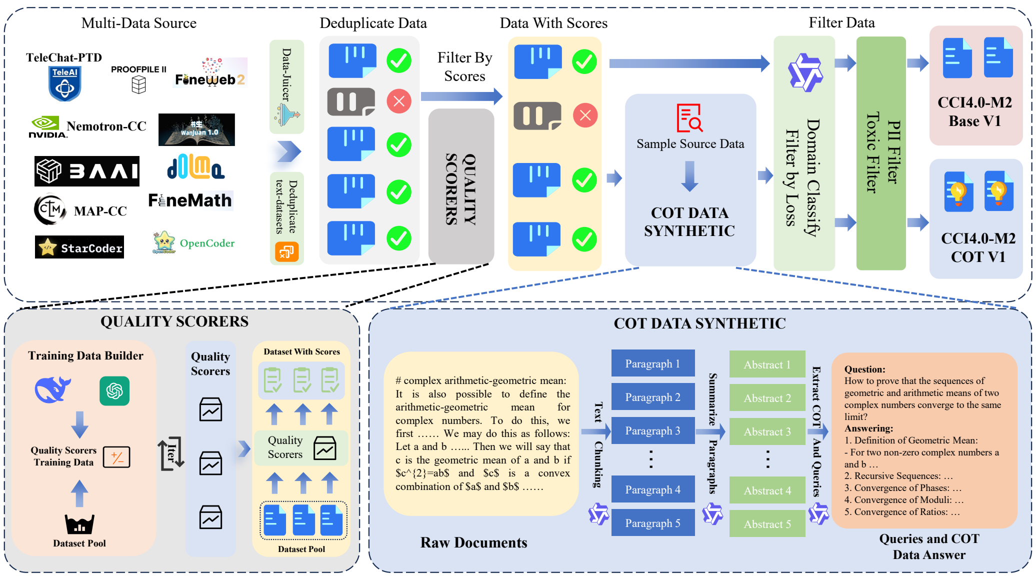

Figure 1: 该图展示了我们数据集的整体处理流程。我们的主要处理流程包括去重、对数据打分以及其他一些筛选过程，以获取我们的 CCI4.0-M2-Base V1 数据集。值得注意的是，我们通过利用大型语言模型进行分块、摘要和提取操作，将从我们数据集池中抽样的原始文档生成 COT 合成数据，以获得 CCI4.0-M2-COT V1 数据集。

As shown in Figure 1, our data processing pipeline is meticulously designed to yield a high-quality, diverse, and robust dataset, comprising five principal stages: Deduplication , Multi-classifier Quality Scoring , Fluency Filtering , Data Synthesis (including CoT synthesis) , and comprehensive Privacy and Toxicity Handling . The initial Deduplication phase is critical for removing redundancy, operating at both a global document level and a finer-grained string level. Following this, the Multi-classifier Quality Scoring stage evaluates data integrity and relevance across various dimensions. This is achieved by automatically constructing evaluation samples using large language models, training specialized small-scale quality classifiers on these samples, and then integrating their scores to assign a comprehensive quality tier to each data point. To address linguistic quality, particularly the prevalence of overly short or syntactically challenging samples, a domain-specific Fluency Filtering step is implemented. Recognizing the significant variability in fluency characteristics across different data domains, this filtering process is applied independently to each domain to effectively remove samples with notably poor linguistic flow. Building upon the foundation of high-quality data identified through the previous stages, a Data Synthesis process is employed. This involves leveraging the filtered high-quality samples as seeds with large models to generate novel data instances in diverse formats. Specifically, high-quality sources are selected for targeted Chain-of-Thought (CoT) synthesis, focusing on the construction of core questions and detailed instructions. Finally, the pipeline incorporates essential safety and privacy measures: Privacy Handling processes sensitive personal information such as identification numbers and phone numbers, while Toxicity Scoring utilizes a dedicated model to assess and flag potentially harmful content within each sample.

3.1 数据收集与预处理

As shown in Table 3, we select multiple sources from English and Chinese. Regarding data sources, the English web corpus was derived from the Nemotron-CC [34] dataset. This particular source was selected based on both comparative effectiveness evaluations and its significantly larger data volume compared to alternatives. For the Chinese web dataset, our collection process involved consolidating data from some existing open-source Chinese datasets such as [4, 5, 6, 10] and extracting the Chinese components from various multilingual datasets such as [17, 26, 27]. Through a thorough analysis of the inter-dependencies and relationships among these potential sources, we strategically filtered and identified over ten datasets that served as the primary contributors to our Chinese web corpus. Furthermore, to enhance the breadth and depth of the training data, we incorporated additional high-performing open-source datasets. These supplementary sources are designed to cover a diverse array of domains, including but not limited to code, mathematics, books, encyclopedias, and academic papers.

Upon conducting manual spot checks on data samples procured from various sources, we identified specific quality inconsistencies, particularly within the Chinese text and code corpora. Consequently, distinct processing methodologies were developed and applied specifically to address these observed issues. For the Chinese dataset, a series of pre-processing operations were undertaken to ensure corpus quality and optimize its suitability for downstream model training. First, to standardize the linguistic and symbolic representation, all text was uniformly converted to Simplified Chinese. Second, to uphold content integrity and compliance with usage norms, a sensitive word filtering mechanism was implemented to automatically detect and remove segments containing inappropriate vocabulary. Furthermore, to mitigate the risk of the model learning overly short or structurally fragmented sentences, a minimum average line length constraint was imposed, retaining only text samples with an average character count of at least 10 per line. Finally, a filter based on the total character count was applied, limiting samples to a range between 100 and 20,000 characters to balance semantic richness with processing efficiency. In parallel, during the processing of the raw code data, we noted the presence of a significant amount of interspersed copyright declarations and related textual information. To construct a high-quality dataset containing solely code content, this non-code material was systematically filtered and removed.

3.2 混合重复数据删除

Following the initial phase of basic quality filtering, a two-stage deduplication strategy was implemented to further enhance the purity and uniqueness of the dataset. The first stage employed a fuzzy deduplication approach[8], leveraging the fuzzy deduplication operator available within the Data-Juicer framework[9]. This method is adept at identifying and eliminating redundant samples by effectively recognizing pairs of texts that are similar in content but not strictly identical. Subsequently, the second stage utilized the deduplicate-text-datasets library[28], an open-source tool developed by Google, to perform exact substring deduplication[21]. This process further removes duplicate data based on precise substring matching. Specifically, we configured the parameters with a length-threshold of 800 and min-doc-words of 35. These settings were carefully chosen to ensure that strict comparisons were performed only between samples exceeding a certain threshold of text length and word count, thereby preventing excessive deduplication of shorter texts. These two complementary deduplication methods work in concert to significantly reduce redundancy while concurrently preserving the diversity of the dataset.

3.3 质量分类

To ensure the high quality of our processed datasets, a multi-faceted quality classification approach was employed, tailored to the characteristics of both English and Chinese corpora. For the English web data, primarily sourced from Nemotron-CC, three independent quality classifiers were utilized to score each document. Based on these scores, samples were allocated into 20 distinct quality bins, with the highest score among the three classifiers designated as the final quality score for each document.

For the Chinese dataset, recognizing the unique linguistic features and the need for domain-specific evaluation, we meticulously designed and trained specialized Chinese quality classifiers during the data construction process. This classification system was conceptually informed by the Nemotron-

CC quality classification framework but underwent significant customization and optimization to effectively handle Chinese linguistic nuances and corpus characteristics. The development of the Chinese quality classifiers involved several critical steps[32]. Initially, we devised specific prompts to guide large models in generating the necessary training data for the classifiers. The training sets were generated using two distinct large models, Qwen2.5-72B-Instruct [35] and Deepseek-V3 [13], yielding a total of 460k samples. The test set, comprising 40k samples, was generated by GPT-4o. The initial sample pools contained 1.2M samples for the training set and 116k samples for the test set. In terms of prompt design, Qwen2.5-72B-Instruct utilized a direct scoring prompt in Chinese, while Deepseek-V3 employed a rule-based cumulative scoring prompt in English. This strategic choice aimed to leverage the respective generative strengths of the two models under different prompting styles.

During the training phase, we fine-tuned two independent XLRoberta-based models[12] on the respective training sets, resulting in two distinct Chinese quality classifiers. We explored four different learning rates (6e-4, 3e-4, 1e-4, and 6e-5) for each configuration, training a complete model under each setting. Evaluation on the test set revealed that when using each classifier independently, a learning rate of 3e-4 yielded the highest F1 score. Furthermore, we observed a significant improvement in the F1 score when combining the outputs of the two classifiers, indicating the complementary nature of the features captured by the data generated from the two different training sets, thereby enhancing the discriminative power of the combined classification system.

In addition to the XLRoberta-based classifiers, and drawing inspiration from the findings presented in A.5, where fastText-based filtering demonstrated optimal performance compared to various model-based data filtering strategies, we also trained a fastText[20] classifier specifically for the Chinese corpus scenario. This classifier was designed as a binary classification task to identify high-quality samples. The positive sample pool was initially constructed by collecting multiple Chinese instruction datasets, including COIG-CQIA[2], OpenHermes-2.5-zh[25], OpenOrca-Chinese and smoltalk Chinese[38]. Through multiple iterations, we progressively refined this set by mitigating the influence of training data length distribution, removing irrelevant high-frequency words from predicted positive samples, and adding certain stop words based on word importance features. This iterative process resulted in a final positive sample set of 220k samples. Correspondingly, 220k samples were randomly drawn from the corpus pool to serve as negative samples. These were then used to construct the final training set of 400k samples and a test set of 40k samples, maintaining a 10:1 ratio.

3.4 基于 LLM 的流畅性过滤

To further refine data quality based on linguistic fluency as [16], we employed a multilingual domain classifier [2] to categorize all raw corpus data into distinct domains, resulting in 26 identified sub-domains. Following this domain classification, we computed the Perplexity Loss for all samples within each domain. The analysis of these loss distributions revealed significant variations across different domains. Notably, the 'Games' domain exhibited the highest overall loss values, suggesting that domains such as gaming contain more complex, unpredictable, or specialized linguistic patterns. Conversely, domains like 'Law and Government' and 'Science and Health' showed the lowest average loss, indicating the presence of more established terminology and formal structures. To mitigate the influence of extreme outliers within each domain, we established a filtering criterion based on the calculated loss distributions. Specifically, samples exceeding the 99.5_{th} percentile of the loss value within their respective domains were systematically removed, yielding a refined dataset with improved linguistic consistency within each category. Loss and percentiles across domains are illustrated in the Figure 5 in Appendix.

3.5 数据合成

Recent studies indicate that the reasoning abilities of large language models (LLMs) primarily originate from the pre-training phase and are subsequently activated during the reinforcement learning phase[15, 39]. Consequently, we endeavor to extract vast quantities of high-quality human thought processes from pre-training corpora to synthesize reasoning data. Specifically, we first curated high-

[2] `https://huggingface.co/nvidia/multilingual-domain-classifier`

quality source data from diverse domains, including web pages, code, mathematics, academic papers, and encyclopedias. As illustrated in the Figure 1, our detailed methodology is as follows:

- 语义分割和摘要：我们利用 Qwen2.5-32B-Instruct [35] 对原始文本进行语义分割。这个过程将原始文档划分为语义上独立且不重叠的片段。为了最小化 LLM 的输出成本，输出仅包含每个片段的起始和结束标记。随后，对于每个片段，我们提示模型生成简洁的摘要。
- 总结思维链和核心问题：基于分段的总结，我们已经能够重建原始人工编写文档背后的思维过程。然后，我们使用 Qwen2.5-32B-instruct 来完善和整合这些分段总结，从而形成一个逻辑连贯的思维链（CoT）。鉴于近期的研究强调问题是推理数据的关键元素，我们最终根据这个思维链推导并总结出文档所涉及的核心问题。

Consequently, each synthesized reasoning data instance is structured as: { core question, chain-of-thought, original document } . In total, we have synthesized over 400 billion (400B) tokens of reasoning data spanning these diverse domains, including web pages, code, mathematics, academic papers, and encyclopedias.

4 实验设置

4.1 训练配置

Following the training setup of CCI3-HQ[36], we adopt the Qwen2-0.5B[37] tokenizer and model architecture, training on a bilingual dataset containing 100 billion tokens. This setup is designed to effectively accommodate both Chinese and English data while preserving consistency across experiments. The training is conducted with a sequence length of 4096, weight decay set to 0.1, and gradient clipping at 1.0. The dataset consists of 25 million samples, trained with a global batch size of 1024. The learning rate follows a cosine decay schedule, starting at 3e-4, decaying to a minimum of 3e-5, with a warmup over the first 2,048,000 samples.

4.2 评估指标

We used the LightEval[14] library for model evaluation, following the same setup as in FineWeb[33] and CCI3-HQ. All evaluations were conducted in a zero-shot setting. To directly compare the performance across different datasets, we use Average, which refers to the overall average score across all Chinese and English benchmarks. The evaluation metrics include:

- 中文基准：CEval [19] 和 CMMLU [23] 。
- 英语基准：ARC-C [11] 、ARC-E [11] 、HellaSwag [40] 、WinoGrande [22] 、MMLU [18] 、OpenbookQA [3] 、PIQA [7] 和 SIQA [29] 。

5 实验结果

5.1 主要结果

We perform isolated training runs using individual datasets to enable direct comparisons between different datasets, such as Nemotron-CC-HQ(the high-quality subset of Nemotron-CC), CCI3-HQ(the high-quality subset of CCI3), and the whole CCI4.0 dataset introduced in this work. The Figure 2 clearly demonstrates that across varying training scales (measured in tokens), the CCI4.0 dataset consistently outperforms both CCI3-HQ and Nemotron-CC-HQ, indicating superior training efficiency and better generalization performance. Key experimental findings are listed as follows:

- 在小规模训练（$\leq$ 20B 个标记）中：CCI4.0 表现出显著的性能优势，特别是在 10B 和 20B 标记时，它以显著的幅度超越其他数据集。这表明数据质量和信息密度较高。例如，CCI4.0 在 10B 标记的表现可以与 Nemotron-CC-HQ 在 30B 标记时的表现相媲美，突显其高效性。
- 在中等规模训练（20B-60B tokens）时：CCI4.0 继续稳步提升，且比其他数据集更早到达性能饱和，这表明在有限的计算预算下，它使模型更快速地接近性能上限。

- 在大规模训练（> 60B 个 tokens）时：虽然不同数据集之间的性能差距缩小，但 CCI4.0 保持领先，表明即使在更大规模的训练下，它仍然有效，并且没有出现早期过拟合或收益递减的情况。
- 总体趋势：CCI4.0 在所有训练阶段始终获得最高分，并表现出稳定的性能曲线，这表明了强大的泛化能力和训练的稳定性。

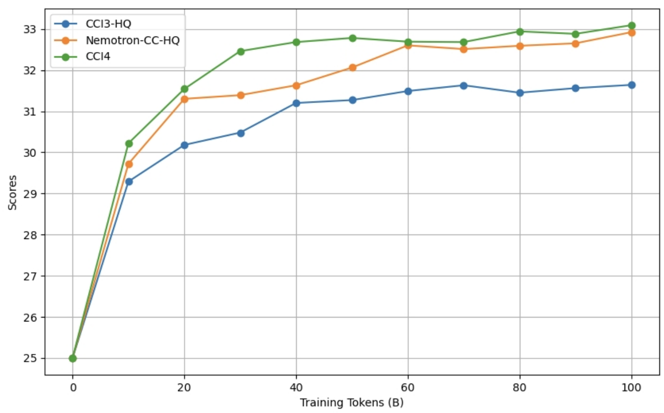

Figure 2: 不同数据集在训练规模上的性能比较。

Table 2: 不同数据集在各个基准测试中的评估结果。

Metrics/Datasets	CCI3-HQ	Nemotron-CC-HQ	CCI4.0
HellaSwag	28.06	44.63	42.50
ARC (Average)	31.03	43.21	41.05
PIQA	55.66	69.15	68.77
MMLU (cloze)	26.52	30.32	30.34
CommonsenseQA	21.05	27.19	27.44
TriviaQA	1.28	5.91	6.05
WinoGrande	48.62	51.38	51.46
OpenBookQA	25.20	33.40	32.60
SIQA	40.43	41.76	40.79
CEval	32.31	27.74	27.67
CMMLU	32.51	26.84	28.92
Average $_{English}$	30.87	38.55	37.89
Average $_{Chinese}$	32.41	27.29	28.30
Average	31.64	32.92	33.09

The Table 2 provides a comprehensive comparison of model performance across multiple benchmarks using three datasets: CCI3-HQ, Nemotron-CC-HQ, and CCI4.0. Although Nemotron-CC-HQ slightly outperforms CCI4.0 on the English average (38.55 vs. 37.89), CCI4.0 remains highly competitive and achieves the best score on several tasks such as CommonsenseQA (27.44) and TriviaQA (6.05). It also matches or closely trails Nemotron on others like MMLU (30.34 vs. 30.32), OpenBookQA (32.60 vs. 33.40), and SIQA (40.79 vs. 41.76), demonstrating robustness across diverse reasoning and knowledge benchmarks. One of CCI4.0’s most notable strengths lies in Chinese language tasks. It significantly outperforms Nemotron-CC-HQ in both CEval (27.67 vs. 27.74) and CMMLU (28.92 vs. 26.84), leading to a higher average Chinese score (28.30 vs. 27.29). This suggests that CCI4.0 contains a more representative and better-curated set of Chinese training data, leading to improved bilingual capabilities. Compared to CCI3-HQ, CCI4.0 shows major improvements in English benchmarks (especially ARC, MMLU, CommonsenseQA, and OpenBookQA), while exhibiting a slight disadvantage in Chinese performance. This is mainly due to the fact that

Chinese data accounts for only around 20 % of the overall dataset composition. We aim for the CCI4.0 dataset to achieve a balanced trade-off between Chinese and English, making it a strong candidate for high-quality pretraining.

5.2 来自 CoT 数据集的推理能力

To evaluate the impact of our CoT dataset on model reasoning abilities, we conducted a controlled experiment. Our evaluation approach is inspired by the adversarial CoT framework proposed by [1], which assesses model reflection on reasoning chains. Given the relatively small scale of our evaluated model, which does not exhibit emergent CoT generation capabilities, directly applying standard CoT evaluation methods is challenging. Therefore, we adapted the evaluation approach. For each test sample containing both a correct and an incorrect CoT, we measure the model's perplexity (PPL) on both CoTs. A sample is considered passed if the model assigns a lower PPL to the correct CoT compared to the incorrect one. The final score for a dataset is the proportion of samples that passed this PPL criterion. Following [1], we report the pre-training compute for each data point as $6nT$, where n and T are the number of parameters and training tokens, respectively.

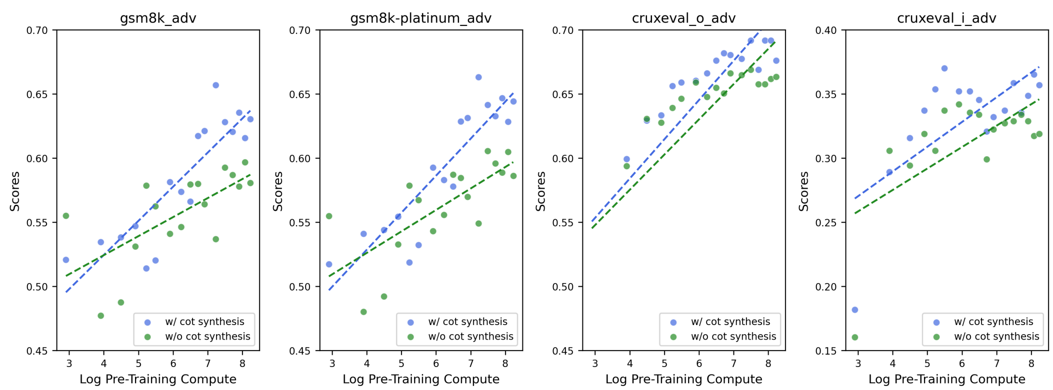

Figure 3: 在对抗数据集上对一个训练了 0.5B 密集模型的推理能力得分进行评估，此模型分别在有（使用 cot 合成）和没有（不使用 cot 合成）CoT 数据混合的情况下接受训练，并在 100B 训练标记上进行评估。使用 CoT 数据进行训练加速了推理能力的增长。

We trained a 0.5B parameter dense language model on two distinct 100 billion-token datasets: one dataset included a mix of CoT data, while the other did not. We evaluated the reasoning performance of models checkpoints using our adapted PPL method on four adversarial datasets from gsm8k_adv, gsm8k-platinum_adv, cruxeval_o_adv, and cruxeval_i_adv. As shown in Figure 3, compared to the model trained without CoT data, the model trained on the dataset mixed with CoT data demonstrates lower perplexity on correct CoT examples, indicating a notably faster improvement in reasoning ability across the evaluated datasets. This demonstrates that incorporating our CoT examples into the training data significantly reduces the model's tendency to hallucinate incorrect CoT examples, accelerating the acquisition of reasoning skills even in smaller models. Further experiments presented in Appendix A.3 provide additional evidence that reasoning ability generally increases with the total training compute.

5.3 从 CoT 数据集得到的下游任务表现

To analyze the influence of CoT Datasets on the model performance, we provide the average model performance across downstream tasks in Figure 4, where models are trained using 10 billion-token datasets with and without CoT data. More detailed model performance is provided in Table 4. Results demonstrate that our synthetic CoT data contributes to performance gains in downstream tasks during model pretraining. Specifically, as shown in Table 4, the model trained with CoT data performs well in reasoning tasks like HellaSwag and reading comprehension tasks like TriviaQA. However, the performance gains brought by CoT data to pretrained models on downstream reasoning tasks are inconsistent, and how to better leverage the effects of CoT data introduced during pretraining in the post-training stage warrants further investigation.

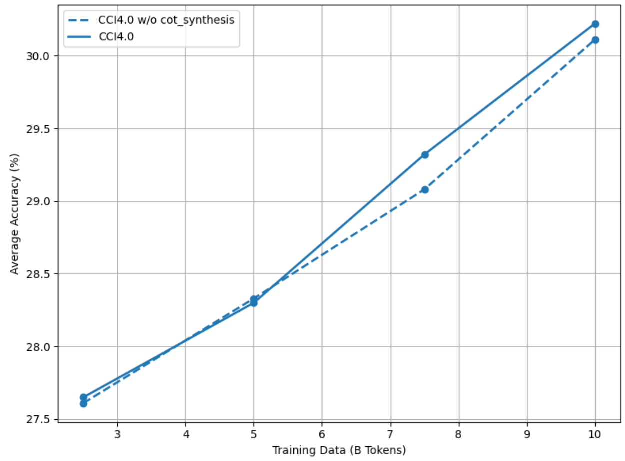

Figure 4: 具有和不具有 CoT 合成的模型性能比较

6 结论

In this work, we introduced CCI4.0, a large-scale, bilingual pretraining dataset designed to provide high-quality and diverse coppora for model training. By integrating diverse, high-quality data sources, including Nemotron-CC for English and multiple Chinese datasets, alongside 4.5 billion human thinking templates via CCI4.0-M2-CoT, CCI4.0 addresses the limitations of traditional datasets in fostering general and complex reasoning abilities. The dataset’s rigorous processing pipeline—encompassing deduplication, multi-classifier quality scoring, fluency filtering, CoT synthesis, and privacy/toxicity handling—ensures both quality and diversity. Experimental results demonstrate that models pretrained on CCI4.0 significantly outperform baselines on reasoning-intensive benchmarks like MMLU and ARC-Challenge with a lower possibility of hallucination. These findings underscore the value of high-quality and diverse pretraining data and establish CCI4.0 as a new standard for developing LLMs capable of tackling sophisticated, multi-step reasoning challenges. Future work will explore further scaling and refinement of CoT synthesis to unlock even greater reasoning potential in next-generation models.

7

致谢

We gratefully acknowledge the valuable contributions of Institutions Alibaba Cloud (阿里云), Shanghai AI Laboratory (上海人工智能实验室), Huawei (华为), Mobvoi (出门问问), Kingsoft Office Software (金山办公), Kunlun (昆仑万维), ModelBest (面壁智能), Qihoo (奇虎科技), Meituan (美团), MiniMax (稀宇科技), Moonshot AI (月之暗面), Zidong Taichu (紫东太初), Wenge (中科闻歌) and iFLYTEK (科大讯飞) in providing the Chinese data.

References

[1] Essential AI, :, Darsh J Shah, Peter Rushton, Somanshu Singla, Mohit Parmar, Kurt Smith, Yash Vanjani, Ashish Vaswani, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Anthony Polloreno, Ashish Tanwer, Burhan Drak Sibai, Divya S Mansingka, Divya Shivaprasad, Ishaan Shah, Karl Stratos, Khoi Nguyen, Michael Callahan, Michael Pust, Mrinal Iyer, Philip Monk, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, and Tim Romanski. Rethinking reflection in pre-training, 2025.

[2] Yuelin Bai, Xinrun Du, Yiming Liang, Yonggang Jin, Ziqiang Liu, Junting Zhou, Tianyu Zheng, Xincheng Zhang, Nuo Ma, Zekun Moore Wang, Ruibin Yuan, Haihong Wu, Hongquan Lin, Wenhao Huang, Jiajun Zhang, Wenhu Chen, Chenghua Lin, Jie Fu, Min Yang, Shiwen Ni,

and Ge Zhang. Coig-cqia: Quality is all you need for chinese instruction fine-tuning. *ArXiv*, abs/2403.18058, 2024.

[3] Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. Careful selection of knowledge to solve open book question answering. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Beijing Academy of Artificial Intelligence. CCI-Data [Data set]. `https://huggingface.co/datasets/BAAI/CCI-Data`.

[5] Beijing Academy of Artificial Intelligence. CCI2-Data [Data set]. `https://huggingface.co/datasets/BAAI/CCI2-Data`.

[6] Beijing Academy of Artificial Intelligence. WuDaoCorporaText [Data set]. `https://data.baai.ac.cn/datadetail/WuDaoCorporaText`.

[7] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019.

[8] Andrei Z. Broder. On the resemblance and containment of documents. *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29, 1997.

[9] Daoyuan Chen, Yilun Huang, Zhijian Ma, Hesen Chen, Xuchen Pan, Ce Ge, Dawei Gao, Yuexiang Xie, Zhaoyang Liu, Jinyang Gao, Yaliang Li, Bolin Ding, and Jingren Zhou. Data-juicer: A one-stop data processing system for large language models. *Companion of the 2024 International Conference on Management of Data*, 2023.

[10] 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, 2023.

[11] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018.

[12] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *ArXiv*, abs/1911.02116, 2019.

[13] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang,

Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025.

[14] Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight framework for llm evaluation, 2023.

[15] Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

[16] Leo Gao, Stella Biderman, Sid Black, Chris Callison-Burch, Laurence Cohen, Esin Durmus, Ethan Fenoglio, Josh Firestone, Jordan Foster, Sam Gehman, Shachar Gretz, Kristen Hallahan, Dieuwke Hupkes, Nathan Lambert, Ron Le Bras, Zachary Levonian, Luca Lisi, Annika McMillan-Major, Todor Mihaylov, Sewon Min, Colin Raffel, Melissa Roemmele, Baptiste Roziere, Maarten Sap, Vered Shwartz, Daniel Sileo, Sabari Subramanian, Leo Sutawika, Yoav Tewel, Josh Tow, Ethan Tseng, Luke van der Poel, Vinay Venkatakrishnan, Benjamin Wang, Guillaume Wenzek, Thomas Wolf, Yuchen Wu, Yutong Xu, Xi Yang, Michihiro Yasunaga, Peter Yin, Rowan Zellers, Tianwei Zhang, Jun Zhu, Mike Lewis, Federico Petroni, and Aleksandra Piktus. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

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

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[19] 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. In *Advances in Neural Information Processing Systems*, 2023.

[20] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[21] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022.

[22] Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012.

[23] 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, 2024.

[24] Jeffrey Li and Alex Fang etc. Datacomp-lm: In search of the next generation of training sets for language models, 2024.

[25] Wenbo Pan. Openhermes 2.5-zh: A partial chinese translation of openhermes-2.5, 2024.

[26] Guilherme Penedo, Hynek Kydlíek, Vinko Sabolec, Bettina Messmer, Negar Foroutan, Martin Jaggi, Leandro von Werra, and Thomas Wolf. Fineweb2: A sparkling update with 1000s of languages, December 2024.

[27] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alibe, Quentin Chanu, Baptiste Launay, Jean-Baptiste Dehaene, and Hugo Touvron. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[28] Google Research. Deduplicate Text Datasets. `https://github.com/google-research/deduplicate-text-datasets`, 2021. Accessed: 2025-05-16.

[29] Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019.

[30] Luca Soldaini, Rodney Kin Lo, Wajdi Yazdan, Ahmed El-Kishky, Faisal Ladhak, Daniel Murray, Shaked Yom Din, Winston Li, Yingbo Liu, Yanai Elazar, Akshita Bhagia, Dirk Groeneveld, Tim Dettmers, Aleksandra Piktus, Nicola Cancedda, Allie De Lucia, Orr Katz, Leshem Choshen, Qiao Li, Hao Zhao, Avi Wettig, Alexander M Rush, and Samuel A Barnett. Dolma: An open corpus of 3 trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*, 2024.

[31] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2024.

[32] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*, 2024.

[33] HuggingFace Team. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

[34] NVIDIA Team. Nemotron-4 340b technical report. *arXiv preprint arXiv:2412.02595*, 2024.

[35] Qwen Team. Qwen2.5: A party of foundation models, September 2024.

[36] 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, 2024.

[37] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[38] 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, 2025.

[39] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

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

A 附录

A.1 跨域损失值

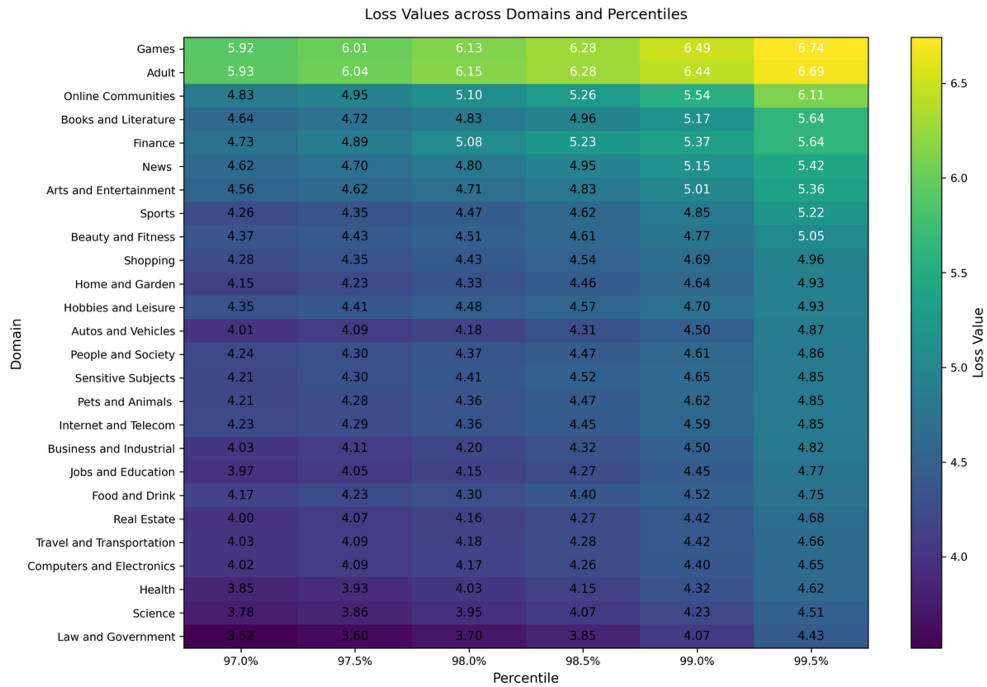

Figure 5: 跨域和百分位数的损失值。

To systematically analyze the model's performance and guide our data filtering strategy, we investigated the distribution of loss values across various domains, as illustrated in Figure 5. The heatmap visualizes the loss values at high percentiles, ranging from the 97th to the 99.5th, offering a granular view of the most challenging instances within the dataset.

A key observation is the significant heterogeneity in loss distribution among the domains. Domains such as "Games" and "Adult" consistently exhibit higher loss values across all percentile thresholds, suggesting they contain a greater concentration of complex, noisy, or out-of-distribution samples that the model struggles to generalize. In contrast, domains like "Law and Government" and "Science" demonstrate substantially lower loss values, indicating a better model fit and cleaner data.

Based on this analysis, we established a data filtering threshold. The decision required a careful trade-off between removing potential noise and preserving valuable information, complicated but valid training examples. A lower percentile threshold would be too aggressive, potentially removing many informative samples. Therefore, we opted to perform the filtering at the 99.5th percentile. This conservative yet precise strategy targets only the most extreme outliers—the top 0.5 % of samples with the highest loss in each domain. This approach allows us to effectively prune the dataset of a majority of probable label errors and severe anomalies while retaining 99.5 % of the data, thus striking an optimal balance between enhancing data quality and maintaining the dataset's scale and diversity for robust model training.

A.2 数据来源

Table 3 provides a comprehensive list of the primary sources considered during our curation process. For the English component, we utilized $Nemotron-CC$, which is derived from Common Crawl. The Chinese component is more extensive, drawing from a variety of prominent web-scale datasets, including $WanJuan$, $WuDaoCorpora$, the $CCI-Dataseries$, and $fineweb-2$, among others. The inclusion of these varied and high-quality sources was crucial for ensuring the breadth, diversity, and scale necessary for our training objectives.

Table 3: 数据集整理过程中考虑的主要数据集

Dataset Type	Dataset Name	URL
Web-EN	Nemotron-CC	https://data.commoncrawl.org/contrib/Nemotron/Nemotron-CC/index.html
Web-ZH	WanJuan	https://opendatalab.org.cn/OpenDataLab/WanJuan1_dot_0
Web-ZH	CCI2.0-Data	https://huggingface.co/datasets/BAAI/CCI2-Data
Web-ZH	CCI1.0-Data	https://huggingface.co/datasets/BAAI/CCI-Data
Web-ZH	WudaoCourpora	https://data.baai.ac.cn/details/WuDaoCorporaText
Web-ZH	ChineseWebText2.0	https://huggingface.co/datasets/CASIA-LM/ChineseWebText2.0
Web-ZH	MAP-CC	https://huggingface.co/datasets/m-a-p/MAP-CC
Web-ZH	TeleChat-PTD	https://huggingface.co/datasets/Tele-AI/TeleChat-PTD
Web-ZH	fineweb-2	https://huggingface.co/datasets/HuggingFaceFW/fineweb-2
Web-ZH	HPLT Datasets v2	https://huggingface.co/datasets/HPLT/HPLT2.0_cleaned
Web-ZH	CCI3.0-Data	https://huggingface.co/datasets/BAAI/CCI3-Data
Code	fineweb-code-corpus_20241112	https://huggingface.co/datasets/OpenCoder-LLM/opc-fineweb-code-corpus
Code	smollm-corpus-python-edu	https://huggingface.co/datasets/HuggingFaceTB/smollm-corpus
Math	fineweb-math-corpus_20241112.jsonl	https://huggingface.co/datasets/OpenCoder-LLM/opc-fineweb-math-corpus
Math	EleutherAI-proof-pile-2-open-web-math.jsonl	https://huggingface.co/datasets/EleutherAI/proof-pile-2
Math	finemath-3plus.jsonl	https://huggingface.co/datasets/HuggingFaceTB/finemath
Books	dolma-books	https://huggingface.co/datasets/allenai/dolma/blob/main/urls/v1_6.txt
Wiki	dolma-wiki	https://huggingface.co/datasets/allenai/dolma/blob/main/urls/v1_6.txt
Arxiv	dolma-arxiv	https://huggingface.co/datasets/allenai/dolma/blob/main/urls/v1_6.txt

Table 3: 数据集整理过程中考虑的主要数据集

Dataset Type	Dataset Name	URL
ForumQA	dolma-v1_7-stackexchange	`https://huggingface.co/datasets/allenai/dolma/blob/main/urls/v1_7.txt`

A.3 推理能力分析

This section presents an experiment analyzing how reasoning ability scales with increasing training compute. We trained a 1.4B parameter MoE model (0.4B active) using 800 billion tokens of our proposed CoT data. We evaluated the model's reasoning performance using the adapted PPL method which is introduced in section 5.2.

As illustrated in Figure 6, the evaluation results demonstrate a clear trend: the model's reasoning ability shows a consistent improvement with increasing training compute on the CoT dataset. This finding suggests that training on a large volume of high-quality CoT data enhances the model's capacity to assign higher probability to correct reasoning paths, even in models where explicit CoT generation is not yet fully emergent.

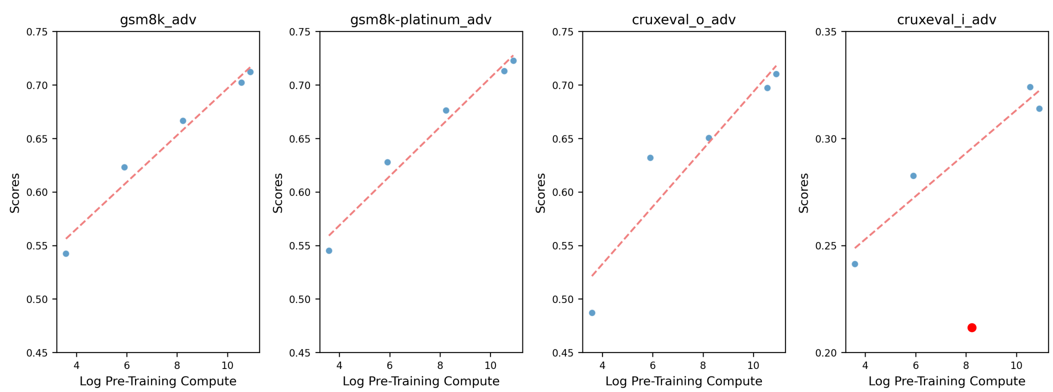

Figure 6: 随着预训练计算量的增加（最多达 8000 亿个 CoT 数据的标记），一个 1.4B MoE 模型在对抗性数据集上的推理能力得分。随着训练计算量的增加，推理能力稳定提高。

A.4 下游任务表现

We provide detailed performance of models trained with and without CoT synthesis dataset in Table 4. Results demonstrate that the model trained with CoT data performs well in reasoning tasks like HellaSwag and reading comprehension tasks like TriviaQA. However, the performance gains brought by CoT data to pretrained models on downstream reasoning tasks are inconsistent, and how to better activate the effects of these CoT data in the post-training stage warrants further investigation.

A.5 中文质量分类器

We apply a combination of our three custom-built Chinese quality classifiers to categorize the Chinese dataset into different quality tiers. Similar to the approach used in Nemotron-CC, we validate the effectiveness of our classifiers by dividing the data into 20 buckets based on quality scores. Separate models are trained and evaluated using data from each bucket. For the Chinese dataset, we focus primarily on the average performance across Chinese evaluation benchmarks.

Table 4: 对比使用和不使用 CoT 综合训练的模型的详细性能。

Metrics/Datasets	CCI4.0 without the CoT Data	CCI4.0
HellaSwag	29.82	30.15
ARC (Average)	33.03	33.16
PIQA	62.79	61.26
MMLU (cloze)	26.62	26.80
CommonsenseQA	25.31	23.67
TriviaQA	0.47	0.79
WinoGrande	49.88	50.28
OpenBookQA	28.00	28.20
SIQA	40.99	40.43
CEval	27.34	27.91
CMMLU	27.11	27.46
Average $_{English}$	32.99	32.75
Average $_{Chinese}$	27.23	27.69
Average	30.11	30.22

The Figure 7 presents the average Chinese evaluation scores of models trained on data buckets categorized by quality scores. Each bucket represents a range of quality scores as assigned by our Chinese data quality classifiers. Several key observations can be made: there is a clear upward trend in model performance as data quality increases. The average Chinese score improves steadily from bucket 0 to bucket 19, indicating that higher-quality data leads to significantly better downstream performance. This validates the effectiveness of the quality classifier in ranking data usefulness.

A.6 基于损失的过滤

To assess the effectiveness of loss-based filtering for English web data, we compare models trained on sampled 10 billion-token English corpora before and after loss filtering. Figure 8 presents the average performance during training. As shown in Figure 8, filtering based on loss improves training efficiency throughout the learning process. Table 5 further indicates that removing outlier samples with high loss from the raw English corpus enhances model performance on downstream commonsense reasoning tasks such as CommonsenseQA and SIQA, as well as reading comprehension tasks like TriviaQA. Notably, although this filtering method is applied solely to English web data, it also leads to slight performance improvements in Chinese QA tasks, such as those in the CMMLU benchmark.

A.7 局限性

While our dataset offers broad coverage and high quality across Chinese and English data, it currently supports only these two languages. Future extensions will aim to incorporate additional languages to better support multilingual modeling and cross-lingual generalization.

Due to the large scale of the dataset, it may not be suitable for researchers with limited computational resources or those working with small models. In such cases, further filtering or subsetting of the dataset may be necessary to ensure practical usability.

Although we have applied privacy-preserving techniques and multiple toxicity filtering strategies using open-source models, we cannot guarantee the complete removal of all sensitive or harmful

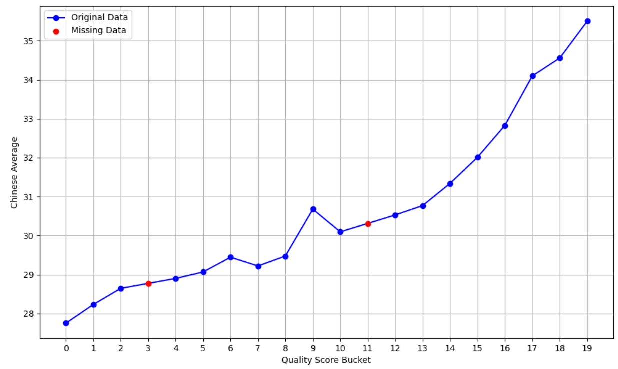

Figure 7: 不同数据集在训练规模上的性能比较。

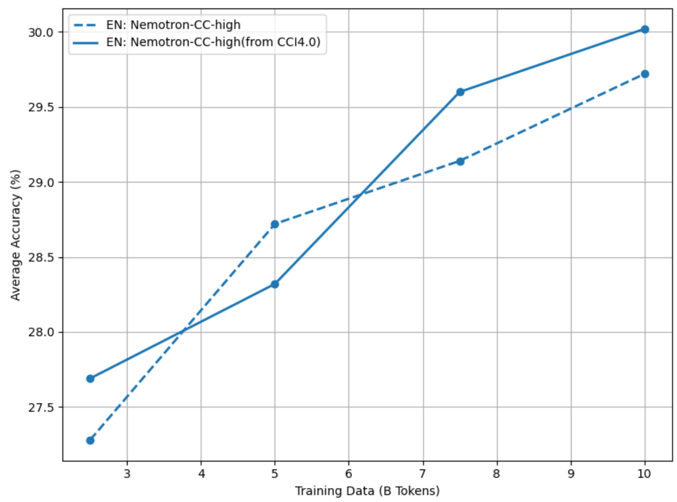

Figure 8: 有无基于损失过滤的模型性能比较

content. Users are advised to apply additional safeguards when deploying models trained on this dataset in sensitive applications.

Table 5: 比较有无损失基础过滤的模型详细性能。

Metrics/Datasets	Nemotron-CC-high	Nemotron-CC-high(from CCI4.0)
HellaSwag	30.93	30.69
ARC (Average)	34.39	34.43
PIQA	62.73	62.89
MMLU (cloze)	26.50	26.64
CommonsenseQA	23.91	25.55
TriviaQA	0.92	1.25
WinoGrande	49.57	49.09
OpenBookQA	29.60	30.20
SIQA	39.56	40.12
CEval	26.65	26.56
CMMLU	25.98	26.67
Average $_{English}$	33.12	33.43
Average $_{Chinese}$	26.32	26.62
Average	29.72	30.02