# AutoMathText-V2:
# A 2.46 Trillion Token AI-Curated STEM Pretraining Dataset

**Chao Li** **Yifan Zhang** **Yang Yuan** **Andrew C Yao**

**Abstract**

We introduce `AutoMathText-V2`, a massive-scale, high-quality pretraining dataset curated for large language models (LLMs) with a strong concentration on Science, Technology, Engineering, and Mathematics (STEM) domains. The dataset consists of **2.46 trillion tokens** derived from over 50 premium data sources, spanning mathematics, code, reasoning, bilingual text, and general web content. To ensure exceptional data quality, we developed a meticulous processing pipeline featuring critical stages: (1) An *three-tier deduplication* process combining exact hash matching, fuzzy deduplication (MinHash+LSH), and advanced semantic deduplication using GTE embeddings. (2) An *AI-powered quality assessment* model, utilizing a fine-tuned Qwen2-based classifier with multi-source score fusion to score and filter content. (3) *Advanced text cleaning* powered by *Ultimate Data Cleaner*, which provides robust, high-performance sanitation while protecting vital STEM content such as complex LaTeX and code blocks. This comprehensive curation process makes `AutoMathText-V2` a superior resource for training robust and capable foundation models.

`Date:` August 20th, 2025

`Website:` https://iiis-ai.github.io/AutoMathText-V2

`Dataset:` https://huggingface.co/datasets/OpenSQZ/AutoMathText-V2

## 1 Introduction

The advancement of Large Language Models (LLMs) is intrinsically linked to the scale and quality of their pretraining data. While general-domain datasets have enabled significant progress, there remains a critical need for high-quality, specialized data in complex domains such as Science, Technology, Engineering, and Mathematics (STEM). STEM fields present unique challenges, including intricate mathematical notation (LaTeX), structured code, logical reasoning, and domain-specific terminology, which are often underrepresented or poorly formatted in general web crawls.

To address this gap, we introduce `AutoMathText-V2` (Li et al., 2025), a massive-scale, 2.46 trillion token pretraining dataset meticulously curated with a strong emphasis on STEM content. Our dataset aggregates over 50 premium data sources and employs a sophisticated, multi-stage processing pipeline to ensure exceptional quality, diversity, and utility. The key contributions of our work are:

- **STEM Concentration:** A purpose-built dataset optimized for mathematics, code, and scientific reasoning to enhance LLM capabilities in technical domains.

- **Three-Tier Deduplication:** An aggressive deduplication strategy combining exact, fuzzy (MinHash-LSH), and semantic (GTE embeddings) methods to maximize data diversity and efficiency.
- **AI-Powered Quality Assessment:** A novel quality scoring system using a fine-tuned Qwen2-based classifier to systematically identify and rank high-quality content.
- **Advanced Text Cleaning:** Robust sanitation using `Ultimate Data Cleaner v7.5.0.5` to normalize text while preserving the integrity of complex structures like LaTeX and code.
- **Contamination Prevention:** Proactive detection and removal of benchmark test questions from math and reasoning datasets to ensure the validity of downstream evaluations.

`AutoMathText-V2` provides the research community with a superior resource for training powerful, robust, and versatile foundation models capable of excelling at complex reasoning and problem-solving tasks.

# 2 Dataset Composition

`AutoMathText-V2` is a comprehensive collection of 2.46 trillion tokens, amalgamating 52 distinct datasets organized into several high-level domains. The distribution is designed to provide a strong foundation in general web text while significantly boosting representation in STEM-focused areas.

## 2.1 Token Distribution by Domain

The dataset is dominated by high-quality web content from Nemotron-CC and DCLM, complemented by substantial portions of code, educational text, reasoning tasks, and specialized mathematics data. Table 1 provides a detailed breakdown of the token count per domain.

## 2.2 Data Sources

The dataset is built from 52 premium sources, each chosen for its quality and relevance. A complete list of sources organized by domain is provided in the Appendix.

# 3 Processing Pipeline

To ensure the highest data quality, every sample in `AutoMathText-V2` was subjected to a rigorous five-stage processing pipeline.

## 3.1 Data Extraction & Standardization

Data from all 52 sources was extracted and standardized into a consistent JSON format. Each entry includes the text content, a unique ID, token count, and metadata such as the original source, domain, and quality scores.

```
{
    "domain_prefix": "lbty.org",
    "id": "117b6a7d-5126-41fe-9bc2-d276e98632e6",
```

**Table 1** Token Distribution by Domain in AutoMathText-V2

| Domain | Token Count (Billions) | Percentage | Description |
|---|---|---|---|
| **Nemotron CC High (Su et al., 2024)** | 1,468.3B | 59.7% | High quality CommonCrawl data |
| **DCLM (Li et al., 2024)** | 314.2B | 12.8% | DCLM baseline web content |
| **RefineCode (Huang et al., 2024)** | 279.4B | 11.4% | GitHub repositories (Academic Use Only) |
| **Nemotron CC Medium-High** | 254.5B | 10.3% | Medium-high quality CommonCrawl data |
| **FineWeb Edu (Penedo et al., 2024)** | 117.4B | 4.8% | Educational web content |
| **Chinese** | 112.18B | 4.6% | Chinese general content |
| **Reasoning QA** | 86.2B | 3.5% | Instruction-following and complex reasoning tasks |
| **Math Web** | 68.3B | 2.8% | Mathematics and scientific content |
| **MegaMath (Zhou et al., 2025)** | 28.5B | 1.2% | Specialized mathematical collections |
| **Translation (Ziemski et al., 2016)** | 1.61B | 0.1% | English-Chinese translation pairs |
| **Total** | **2,460.71B** | **100%** | **Complete dataset** |


```
  "meta": "{\"domain\": \"dclm\", \"ori_score\":
 0.043276190757751465, \"source\": \"dclm_baseline\"}",
  "text": "Sabine Expedition\n\nThe Sabine Expedition was an
 expedition approved by the United States Congress in 1806...",
  "tokens": 145,
  "url": "https://lbty.org/american-indian-battles/sabine-
 expedition/",
  "score": 0.19072403013706207
}
```


**Listing 1** Standardized data format example

### 3.2 Three-Tier Deduplication

We employed a multi-stage deduplication process to maximize data novelty and remove redundant information.

#### 3.2.1 Exact Deduplication

We first performed exact deduplication using SHA256 hashing on the text content. In cases of collision, priority was given to sources deemed higher quality. This initial pass removed approximately 30% of documents.

#### 3.2.2 Fuzzy Deduplication

Next, we applied MinHash Locality Sensitive Hashing (LSH) to identify near-duplicates. Documents were clustered using a Jaccard similarity threshold of 0.9. Within each cluster, only the document with the highest quality score was retained. This stage removed an additional 20% of near-duplicate documents.

#### 3.2.3 Semantic Deduplication

Finally, we performed semantic deduplication to remove documents with similar meaning but different phrasing. We generated embeddings using `Alibaba-NLP/gte-multilingual-base` and used K-means clustering (k=100,000) to group semantically similar documents. A cosine similarity threshold of 0.007 was used to filter duplicates within clusters, removing a final 10% of the data.

### 3.3 AI Quality Assessment

A fine-tuned Qwen2 model was used as a quality classifier. The model was trained with a regression head to predict a quality score for each document (Zhang et al., 2025). Scores from multiple sources were normalized and fused to produce a final, reliable quality metric used for filtering and for creating quality-based percentile splits in the final dataset.

### 3.4 Advanced Text Cleaning

All text was processed with `Ultimate Data Cleaner v7.5.0.5`. This tool was configured for high-performance cleaning of web-scraped and scientific data. Key features included advanced protection for nested LaTeX environments and markdown code fences, alongside quality heuristics to remove corrupted text (e.g., excessive repetition, bracket imbalances).

### 3.5 Contamination Detection

To ensure the integrity of model evaluation, we implemented a strict contamination detection protocol. Test set questions from standard benchmarks (e.g., GSM8K, MATH) were compiled. We performed exact string matching against our dataset, filtering out any documents that contained benchmark questions. This process was integrated directly into the data extraction stage to prevent contamination from entering the pipeline.

## 4 Dataset Structure and Usage

### 4.1 Loading with `datasets`

The dataset is available on the Hugging Face Hub and can be easily loaded using the `datasets` library. The full dataset or specific domains can be loaded in streaming mode to handle its large size.

```
from datasets import load_dataset

# Load the full dataset in streaming mode
dataset = load_dataset("OpenSQZ/AutoMathText-V2", streaming=True)

# Load a specific domain (e.g., math_web)
math_data = load_dataset("OpenSQZ/AutoMathText-V2", name="math_web",
    streaming=True)
```

**Listing 2** Loading the dataset with Hugging Face Datasets

### 4.2 RefineCode Content Download

For the `refinecode` domain, only metadata is provided in the main dataset to reduce storage overhead. The full code content must be downloaded from the Software Heritage S3 bucket using the `blob_id` provided in the metadata. The following script demonstrates this process.

```
import os
import json
import boto3
from smart_open import open
from datasets import load_dataset

# Setup AWS credentials from environment variables
session = boto3.Session(
    aws_access_key_id=os.environ["AWS_ACCESS_KEY_ID"],
    aws_secret_access_key=os.environ["AWS_SECRET_ACCESS_KEY"]
)
s3 = session.client("s3")

def download_code_content(blob_id, src_encoding):
    """Download code content from AWS S3 using blob_id."""
    s3_url = f"s3://softwareheritage/content/{blob_id}"
    try:
        with open(s3_url, "rb", compression=".gz", transport_params
   ={"client": s3}) as fin:
            content = fin.read().decode(src_encoding)
        return {"content": content}
    except Exception as e:
        return {"content": None, "error": str(e)}
```

```python
# Load RefineCode domain metadata
refinecode_data = load_dataset("OpenSQZ/AutoMathText-V2", name="
    refinecode", streaming=True)

# Process each sample to download content
for sample in refinecode_data:
    meta = json.loads(sample["meta"])
    blob_id = meta.get("blob_id")
    src_encoding = meta.get("src_encoding", "utf-8")

    if blob_id:
        code_data = download_code_content(blob_id, src_encoding)
        full_sample = {**sample, "code_content": code_data["content"
    ]}
        print(f"Downloaded content for {sample['id']}")
        # Process the full_sample here
        break # Example stops after one sample
```

**Listing 3** Downloading full code content for the RefineCode domain

This requires the `boto3` and `smart_open` libraries and valid AWS credentials with access to the bucket.

## References


Siming Huang, Tianhao Cheng, Jason Klein Liu, Jiaran Hao, Liuyihan Song, Yang Xu, J Yang, Jiaheng Liu, Chenchen Zhang, Linzheng Chai, et al. Opencoder: The open cookbook for top-tier code large language models. *arXiv preprint arXiv:2411.04905*, 2024.

Chao Li, Yifan Zhang, Yang Yuan, and Andrew C Yao. Automathtext-v2: A 2.46 trillion token ai-curated stem pretraining dataset, 2025. https://huggingface.co/datasets/OpenSQZ/AutoMathText-V2. A 2.46T token multi-domain dataset with fine-grained deduplication and AI-powered quality assessment.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849, 2024.

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.

Yifan Zhang, Yifan Luo, Yang Yuan, and Andrew Chi-Chih Yao. Autonomous data selection with zero-shot generative classifiers for mathematical texts. *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025 Findings)*, 2025.

Fan Zhou, Zengzhi Wang, Nikhil Ranjan, Zhoujun Cheng, Liping Tang, Guowei He, Zhengzhong Liu, and Eric P Xing. Megamath: Pushing the limits of open math corpora. *arXiv preprint arXiv:2504.02807*, 2025.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1561.

# Appendix

## A Complete Data Sources

This section provides a complete list of the 52 premium data sources used in the construction of AutoMathText-V2, organized by their respective domains.

**Table 2** Complete List of Data Sources

| Source | HuggingFace Dataset | Description |
|---|---|---|
| | **DCLM Domain** | |
| DCLM-Baseline | DCLM/dclm-baseline-1.0 | High-quality web content from DCLM |
| | **FineWeb Edu Domain** | |
| FineWeb-Edu | HuggingFaceFW/fineweb-edu | Educational web content (0-5 quality scale) |
| | **FineWeb Edu Chinese Domain** | |
| FineWeb-Edu-Chinese | opencsg/Fineweb-Edu-Chinese -V2.1 | Chinese educational content (3.4-5.0 scale) |
| | **Math Web Domain** | |
| AutoMathText | math-ai/AutoMathText | Math/Code/ArXiv content with lm_q1q2_score |
| FineMath | HuggingFaceTB/finemath | High-quality mathematics content (0-5 scale) |
| Open-Web-Math-Pro | gair-prox/open-web-math-pro | Mathematical web pages |
| InfiMM-WebMath-40B | Infi-MM/InfiMM-WebMath-40B | Multimodal mathematical content |
| | **Nemotron CC Domains** | |
| Nemotron-CC (High) | nvidia/nemotron-cc | High-quality CommonCrawl subset |
| Nemotron-CC (Medium-High) | nvidia/nemotron-cc | Medium-high quality CommonCrawl subset |
| | **RefineCode Domain** | |
| RefineCode | m-a-p/RefineCode | GitHub repositories (Academic Use Only) |
| | **Reasoning QA Domain** | |
| OPC-Annealing-Corpus | OpenCoder-LLM/opc-annealing -corpus | Code training corpus |
| OPC-SFT-Stage1 | OpenCoder-LLM/opc-sft-stage1 | Instruction following data (stage 1) |
| OPC-SFT-Stage2 | OpenCoder-LLM/opc-sft-stage2 | Instruction following data (stage 2) |
| Magpie-Reasoning-V2-250K-CoT-QwQ | Magpie-Align/Magpie-Reasoning-V2... | Chain-of-thought reasoning (QwQ) |
| Magpie-Reasoning-V1-150K-CoT-QwQ | Magpie-Align/Magpie-Reasoning-V1... | Chain-of-thought reasoning (QwQ) |

Continued on next page

**Table 2 – continued from previous page**

| Source | HuggingFace Dataset | Description |
|---|---|---|
| Magpie-Reasoning-V1-150K-CoT-Deepseek | Magpie-Align/Magpie-Reasoning-V1... | Advanced reasoning (DeepSeek-R1) |
| Magpie-Reasoning-V2-250K-CoT-Deepseek | Magpie-Align/Magpie-Reasoning-V2... | Advanced reasoning (DeepSeek-R1) |
| General-Instruction-Augmented-Corpora | instruction-pretrain/general-instruction... | General instruction synthesis |
| FT-Instruction-Synthesizer-Collection | instruction-pretrain/ft-instruction... | Fine-tuning instruction synthesis |
| Code-Feedback-Filtered-Instruction | m-a-p/CodeFeedback-Filtered-Instruction | Code QA with feedback |
| XCoder-80K | banksy235/XCoder-80K | Code instruction data |
| Orca-Math-Word-Problems-200K | microsoft/orca-math-word-problems... | Math word problems |
| Meta-Math-QA | meta-math/MetaMathQA | Mathematical QA dataset |
| Numina-Math-CoT | AI-MO/NuminaMath-CoT | Math chain-of-thought |
| Scale-Quest-Math | dyyyyyyyy/ScaleQuest-Math | Mathematical problem solving |
| Calc-Ape210K | MU-NLPC/Calc-ape210k | Chinese math problems |
| MathInstruct | TIGER-Lab/MathInstruct | Math instruction data |
| MathScaleQA-2M | fdqerq22ds/MathScaleQA-2M | Large-scale math QA |
| Gretel-Math-GSM8K-V1 | gretelai/gretel-math-gsm8k-v1 | GSM8K style problems |
| Open-Math-Instruct-2 | nvidia/OpenMathInstruct-2 | Open math instructions |
| Stack-Math-QA | math-ai/StackMathQA | Stack Exchange math QA |
| OpenR1-Math-220K | open-r1/OpenR1-Math-220k | Advanced math reasoning |
| Natural-Reasoning | facebook/natural_reasoning | Natural language reasoning |
| Math-Code-Instruct | MathLLMs/MathCodeInstruct | Math with code instructions |
| Math-Code-Instruct-Plus | MathLLMs/MathCodeInstruct-Plus | Enhanced math-code instructions |
| Open-Orca | Open-Orca/OpenOrca | General instruction following |
| SlimOrca-Deduped-Cleaned | Open-Orca/slimorca-deduped-cleaned... | Cleaned instruction data |
| Orca-AgentInstruct-1M-V1-Cleaned | mlabonne/orca-agentinstruct-1M-v1... | Agent instruction data |
| FOL-NLI | tasksource/FOL-nli | First-order logic reasoning |
| Infinity-Instruct | BAAI/Infinity-Instruct | Multi-domain instructions |
| Llama-Nemotron-Post-Training-Dataset | nvidia/Llama-Nemotron-Post-Training... | Post-training dataset |
| Codeforces-CoTs | open-r1/codeforces-cots | Competitive programming |
| Reasoning-V1-20M | glaiveai/reasoning-v1-20m | Large-scale reasoning data |
| Lean-STaR-Plus | ScalableMath/Lean-STaR-plus | Lean formal proofs (enhanced) |
| Lean-STaR-Base | ScalableMath/Lean-STaR-base | Lean formal proofs (base) |
| Lean-CoT-Plus | ScalableMath/Lean-CoT-plus | Lean chain-of-thought (enhanced) |
| Lean-CoT-Base | ScalableMath/Lean-CoT-base | Lean chain-of-thought (base) |
| Lean-Github | internlm/Lean-Github | Lean repository code |
| Lean-Workbook | internlm/Lean-Workbook | Lean problem workbook |
| DeepSeek-Prover-V1 | deepseek-ai/DeepSeek-Prover-V1 | Formal proof verification |

Continued on next page

**Table 2 – continued from previous page**

| Source | HuggingFace Dataset | Description |
|---|---|---|
| **Translation Domain** | | |
| UN-PC | Helsinki-NLP/un_pc | English-Chinese translation pairs |
| UN-PC-Reverse | Helsinki-NLP/un_pc | Chinese-English translation pairs |
| **MegaMath Domain** | | |
| MegaMath-QA | LLM360/MegaMath | Large-scale mathematical QA |
| MegaMath-Translated-Code | LLM360/MegaMath | Mathematical code translations |
| MegaMath-Text-Code-Block | LLM360/MegaMath | Mixed math text and code blocks |