Generative AI for Math

MathPile: A Billion-Token-Scale Pre-training Corpus for Math

Zengzhi Wang[1,3,4], Xuefeng Li[1,4], Rui Xia[3], Pengfei Liu[1,2,4]

Presenter: Zengzhi Wang

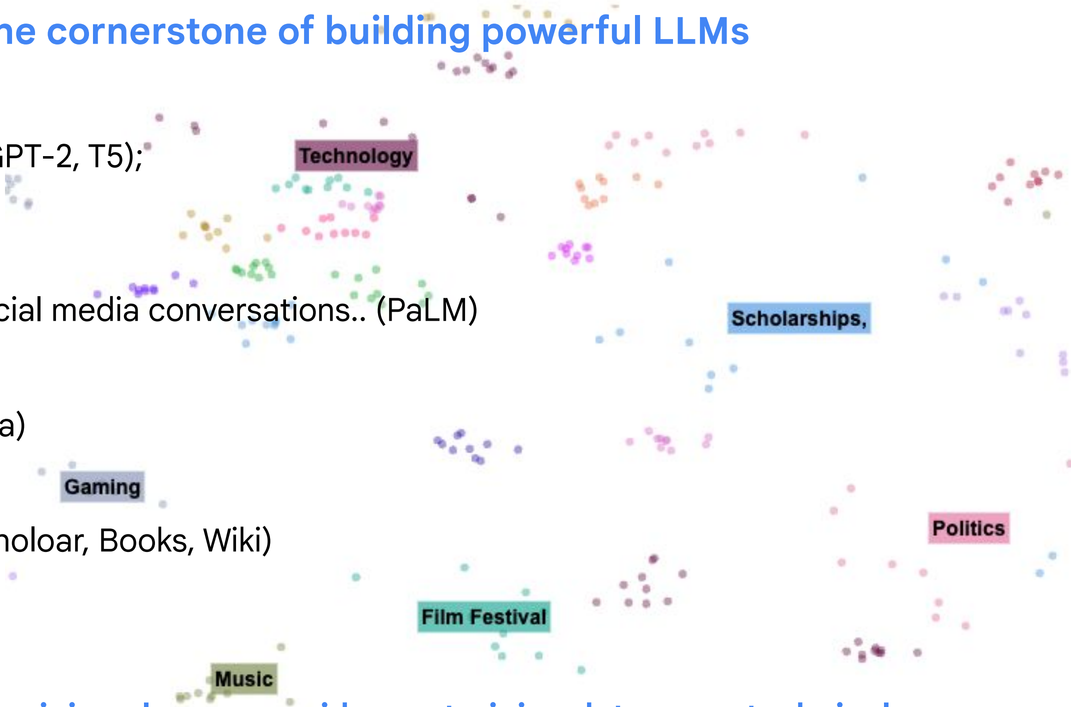
Outline

- Background: Why this dataset?
- How to construct a pre-training corpus for math?
  - data collection, filtering, cleaning, and deduplication
- Experimental Results: demonstrate the effectiveness of corpus and pipeline
- Takeaways

Historical Changes on Pre-training Corpora

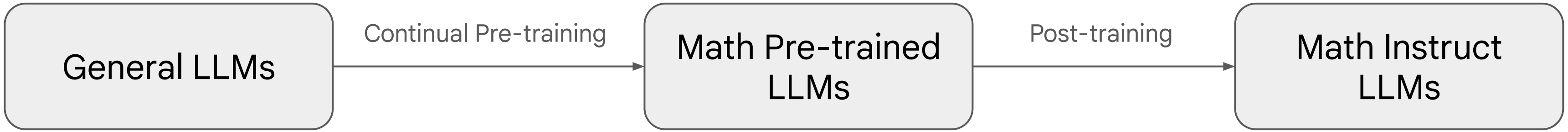
High-quality, large-scale corpora are the cornerstone of building powerful LLMs

- Books and Wikipedia (GPT, BERT);
- Web pages, e.g., reddit and Common Crawl, (GPT-2, T5);
- CC, WebText, Books, Wikipedia (GPT-3)
- Pile (GPT-Neo..)
- MassiveText (Gopher)
- Web pages, Books, Wikipedia, News, Code, Social media conversations.. (PaLM)
- ROOTS (BLOOM)
- The Stack (StarCoder)
- RedPajama (a reproduction of LLaMA’s corpora)
- SlimPajama, RedPajama v2
- RefinedWeb, web-only, (Falcon..)
- Dolma (web pages, code, Reddit, Semantic Scholoar, Books, Wiki)
- FineWeb (-edu) (web-only)
- DCLM-baseline (web-only)



Today, many LLMs, such as GPT-4, Mistral, Gemini, no longer provide pre-training data, even technical details.

Background: Pre-training Corpora for Math



General LLMs

Continual Pre-training

Math Pre-trained LLMs

Post-training

Math Instruct LLMs

Google Minerva's Corpus (38.5B tokens)

OpenAI's MathMix (1.5B tokens)

Unfortunately, closed source

WizardMath (Luo et al., 2023)

Abel (Chern et al., 2023)

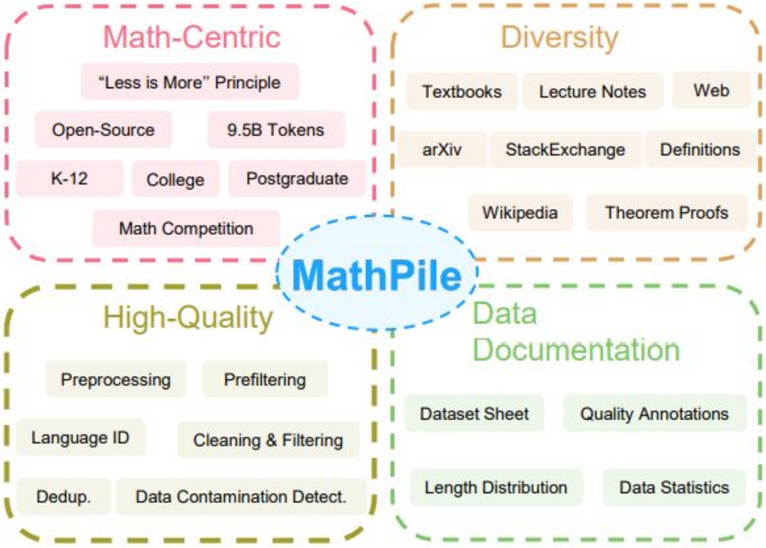
MAmmoTH (Yue et al., 2023)

......

At that time (in 2023), researchers mainly focus on math SFT due to the lack of math pre-training corpora and other factors.

Note that OpenWebMath (Paster et al., 2023) is conducting work concurrently with ours.

MathPile: Overview



MathPile

Math-Centric

"Less is More" Principle

Open-Source

9.5B Tokens

K-12

College

Postgraduate

Math Competition

Diversity

Textbooks

Lecture Notes

Web

arXiv

StackExchange

Definitions

Wikipedia

Theorem Proofs

High-Quality

Preprocessing

Prefiltering

Language ID

Cleaning & Filtering

Dedup.

Data Contamination Detect.

Data Documentation

Dataset Sheet

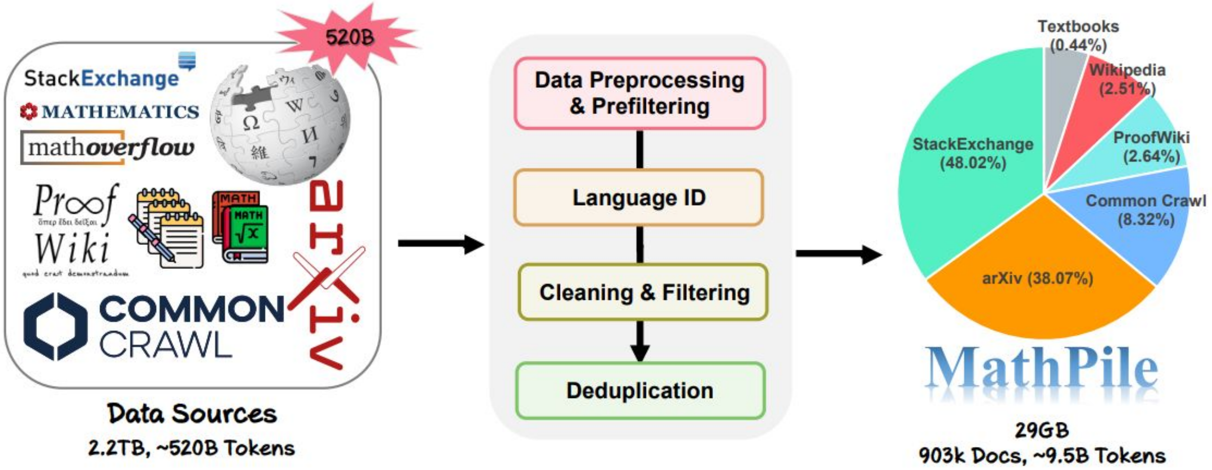
Quality Annotations

Length Distribution

Data Statistics

MathPile: Data Processing Pipeline

- Step 1: Data collection from various sources
  - with necessary preprocessing (cleaning, filtering)
- Step 2: Global Data Processing Pipeline
  - Language Identification, Cleaning & Filtering, Deduplication, Decontamination



MathPile: Statistics

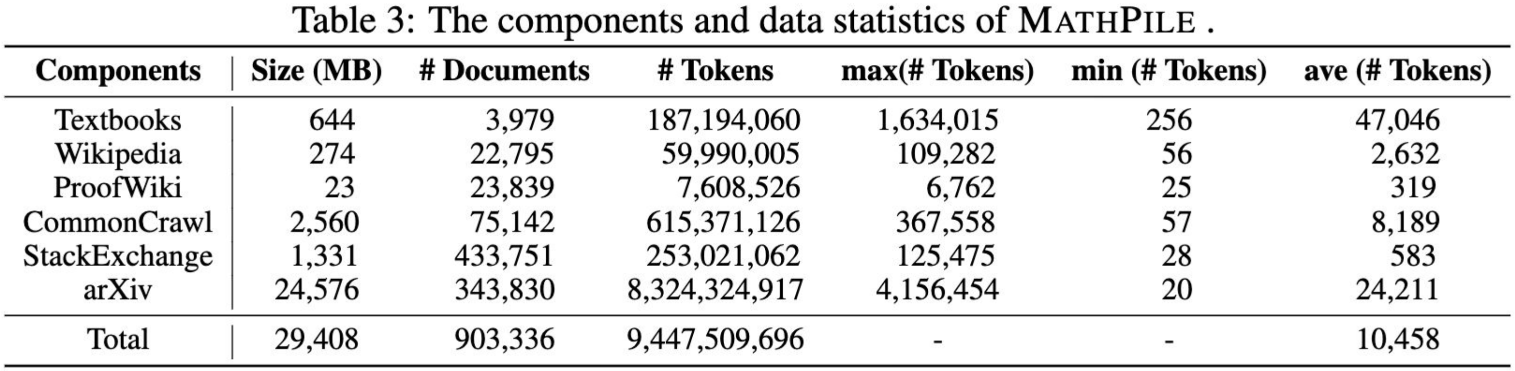


Table 3: The components and data statistics of MATHPILE .

Components	Size (MB)	# Documents	# Tokens	max(# Tokens)	min (# Tokens)	ave (# Tokens)
Textbooks	644	3,979	187,194,060	1,634,015	256	47,046
Wikipedia	274	22,795	59,990,005	109,282	56	2,632
ProofWiki	23	23,839	7,608,526	6,762	25	319
CommonCrawl	2,560	75,142	615,371,126	367,558	57	8,189
StackExchange	1,331	433,751	253,021,062	125,475	28	583
arXiv	24,576	343,830	8,324,324,917	4,156,454	20	24,211
Total	29,408	903,336	9,447,509,696	-	-	10,458

Continual Pre-training Experiments

- Base model: Mistral-7B-v0.1 (SOTA LLM at that time)
- Benchmarks:
    - Elemental Math: GSM8K, MMLU-Math
    - High-school: MATH, AGIEval-SAT-MATH, AQuA, MathQA;
    - College: MMLU-Math
- Few-shot Prompting Evaluation

The Effectiveness of MathPile

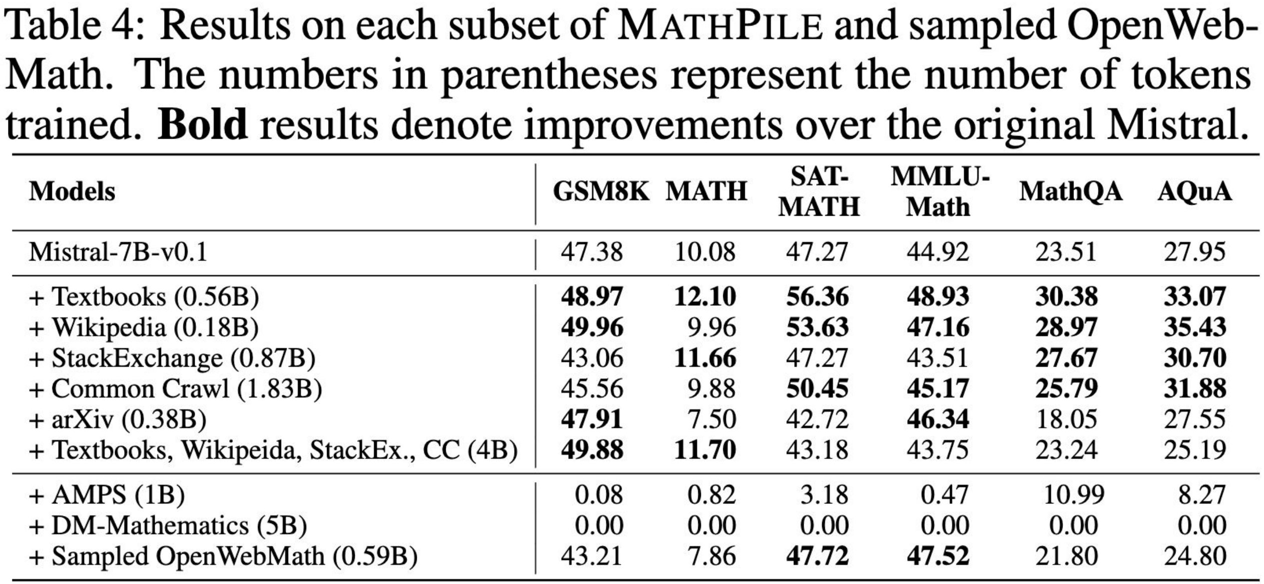
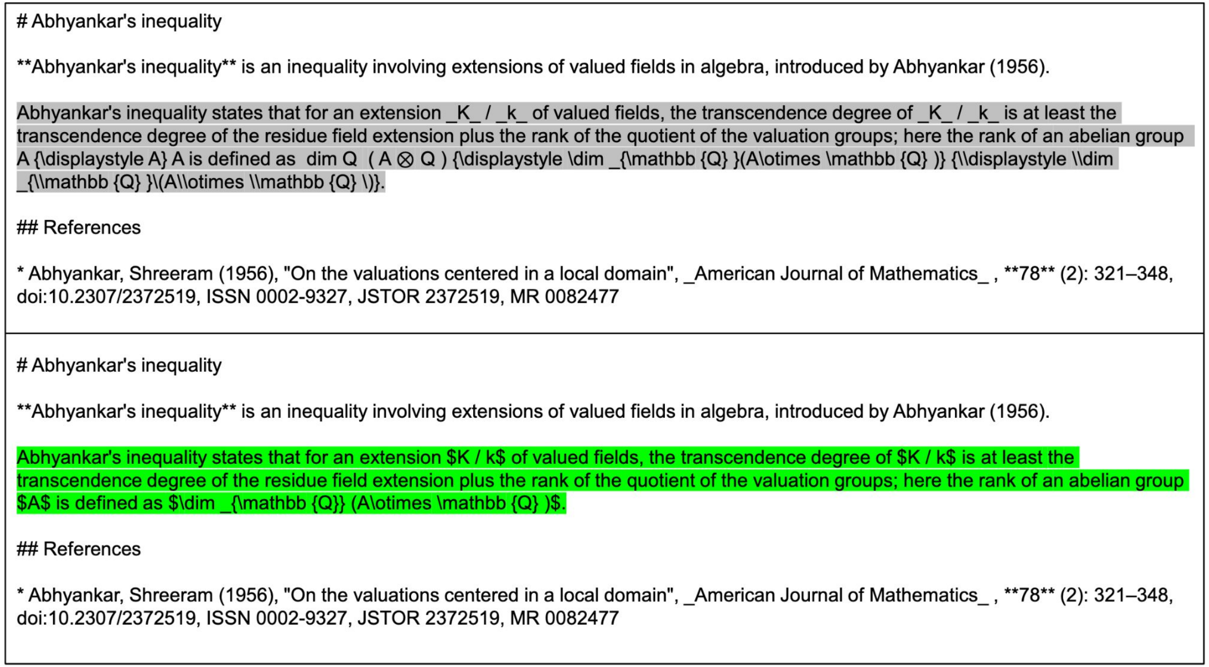


Table 4: Results on each subset of MATHPILE and sampled OpenWebMath. The numbers in parentheses represent the number of tokens trained. **Bold** results denote improvements over the original Mistral.

Models	**GSM8K**	**MATH**	**SAT-MATH**	**MMLU-Math**	**MathQA**	**AQuA**
Mistral-7B-v0.1	47.38	10.08	47.27	44.92	23.51	27.95
+ Textbooks (0.56B)	**48.97**	**12.10**	**56.36**	**48.93**	**30.38**	**33.07**
+ Wikipedia (0.18B)	**49.96**	9.96	**53.63**	**47.16**	**28.97**	**35.43**
+ StackExchange (0.87B)	43.06	**11.66**	47.27	43.51	**27.67**	**30.70**
+ Common Crawl (1.83B)	45.56	9.88	**50.45**	**45.17**	**25.79**	**31.88**
+ arXiv (0.38B)	**47.91**	7.50	42.72	**46.34**	18.05	27.55
+ Textbooks, Wikipeida, StackEx., CC (4B)	**49.88**	**11.70**	43.18	43.75	23.24	25.19
+ AMPS (1B)	0.08	0.82	3.18	0.47	10.99	8.27
+ DM-Mathematics (5B)	0.00	0.00	0.00	0.00	0.00	0.00
+ Sampled OpenWebMath (0.59B)	43.21	7.86	**47.72**	**47.52**	21.80	24.80

The Effectiveness of Data Processing Pipeline

- Taking Wikipedia as an example
- ablation on two decisions:
  - perform
    - spec
  - fix LaTeX



```
# Abhyankar's inequality

**Abhyankar's inequality** is an inequality involving extensions of valued fields in algebra, introduced by Abhyankar (1956).

Abhyankar's inequality states that for an extension _K_ / _k_ of valued fields, the transcendence degree of _K_ / _k_ is at least the transcendence degree of the residue field extension plus the rank of the quotient of the valuation groups; here the rank of an abelian group A {\displaystyle A} A is defined as  dim Q  ( A ⊗ Q ) {\displaystyle \dim _{\mathbb {Q} }(A\otimes \mathbb {Q} )} {\\displaystyle \\dim _{\\mathbb {Q} }\(A\\otimes \\mathbb {Q} \)}.

## References

* Abhyankar, Shreeram (1956), "On the valuations centered in a local domain", _American Journal of Mathematics_ , **78** (2): 321–348, doi:10.2307/2372519, ISSN 0002-9327, JSTOR 2372519, MR 0082477
```

```
# Abhyankar's inequality

**Abhyankar's inequality** is an inequality involving extensions of valued fields in algebra, introduced by Abhyankar (1956).

Abhyankar's inequality states that for an extension $K / k$ of valued fields, the transcendence degree of $K / k$ is at least the transcendence degree of the residue field extension plus the rank of the quotient of the valuation groups; here the rank of an abelian group $A$ is defined as $\dim _{\mathbb {Q}} (A\otimes \mathbb {Q} )$.

## References

* Abhyankar, Shreeram (1956), "On the valuations centered in a local domain", _American Journal of Mathematics_ , **78** (2): 321–348, doi:10.2307/2372519, ISSN 0002-9327, JSTOR 2372519, MR 0082477
```

The Effectiveness of Data Processing Pipeline

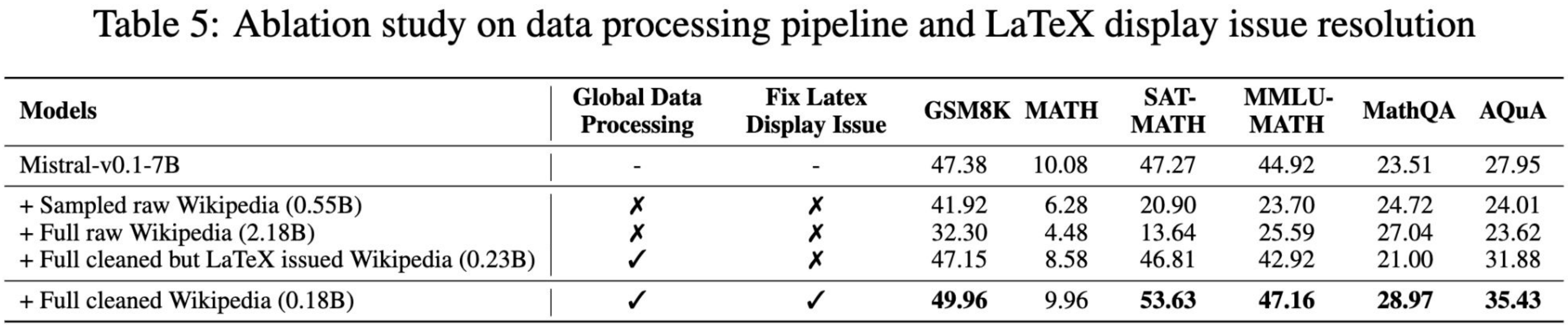


Table 5: Ablation study on data processing pipeline and LaTeX display issue resolution

Models	**Global Data Processing**	**Fix Latex Display Issue**	**GSM8K**	**MATH**	**SAT-MATH**	**MMLU-MATH**	**MathQA**	**AQuA**
Mistral-v0.1-7B	-	-	47.38	10.08	47.27	44.92	23.51	27.95
+ Sampled raw Wikipedia (0.55B)	✗	✗	41.92	6.28	20.90	23.70	24.72	24.01
+ Full raw Wikipedia (2.18B)	✗	✗	32.30	4.48	13.64	25.59	27.04	23.62
+ Full cleaned but LaTeX issued Wikipedia (0.23B)	✓	✗	47.15	8.58	46.81	42.92	21.00	31.88
+ Full cleaned Wikipedia (0.18B)	✓	✓	**49.96**	9.96	**53.63**	**47.16**	**28.97**	**35.43**

Takeaways

- MathPile, a 9.5B tokens corpus for math domain, with diverse sources, covering textbooks, scientific papers, web pages, Community QA, and wiki.
- This corpus have been used in many studies so far, usage including but not limited to pre-training, data synthesis and benchmarking.

- Some limitations:
  - only focus on English
  - many decisions were made empirically, not always optimal.
  - without employing model-based data filtering to improve quality.
  - subset like common crawl could be expanded.

Thanks

- Paper: https://huggingface.co/papers/2312.17120
- Github: https://github.com/GAIR-NLP/MathPile/
- Dataset (Research-only): https://huggingface.co/datasets/GAIR/MathPile
- Dataset (Commerical use): https://huggingface.co/datasets/GAIR/MathPile_Commercial

feel free to email me: zzwang.nlp@gmail.com