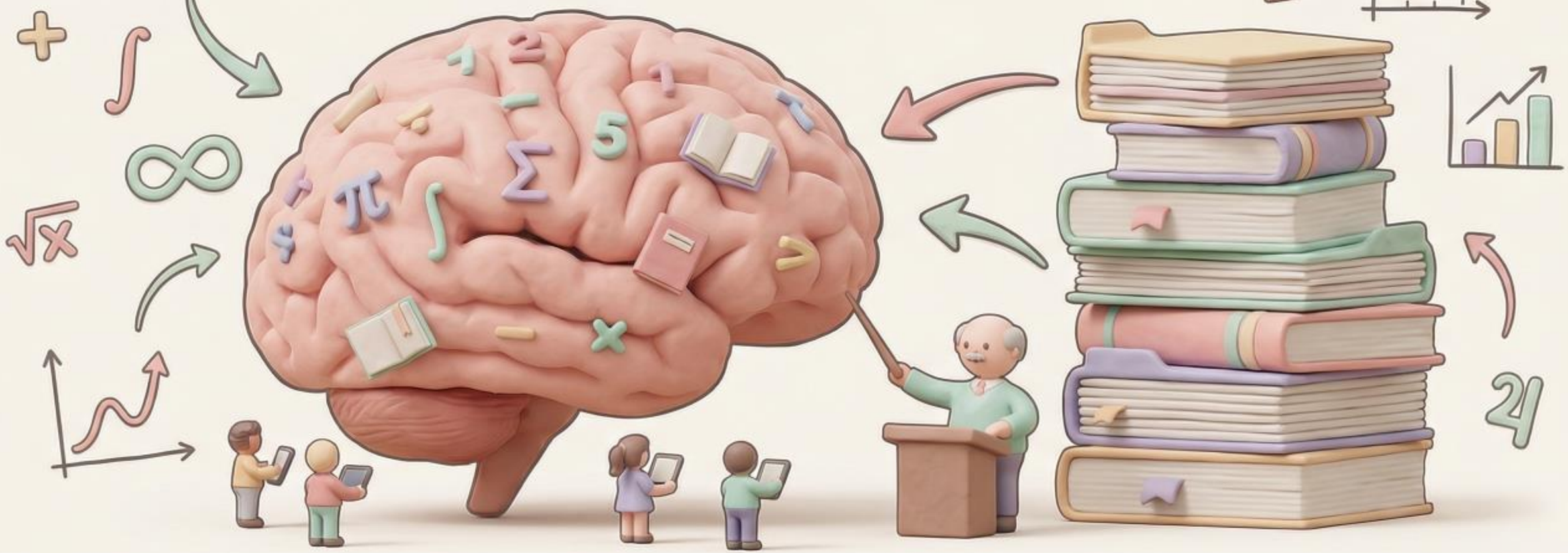
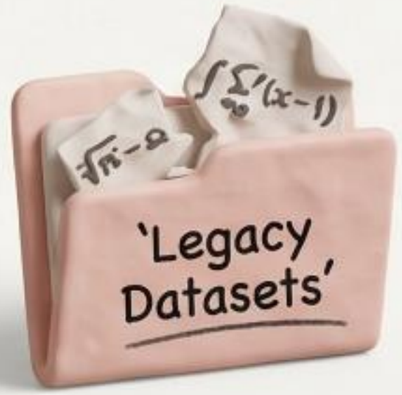
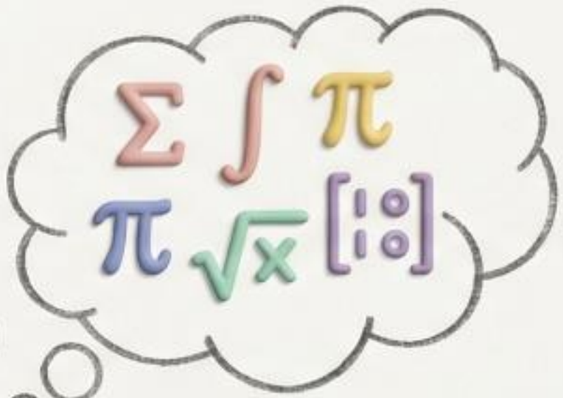
NEMOTRON-CC-MATH

A 133 Billion-Token-Scale High Quality Math Pretraining Dataset



Introduction: The Challenge of Mathematical Data





LLM Reasoning Capabilities

Critical for LLM reasoning, but existing datasets are brittle & lossy.

Proprietary (Minerva, DeepSeekMath)

Quality Gap

Open-Source Alternatives

- Brittle extraction heuristics.
- Lossy HTML-to-text conversion.
- Fails to preserve structure.

Overview of Nemotron-CC-Math

Highest Quality Subset

Nemotron-CC-Math-4+ | 52B Tokens

Processed Data

Nemotron-CC-Math-3+ | 133B Tokens Total

Raw Data

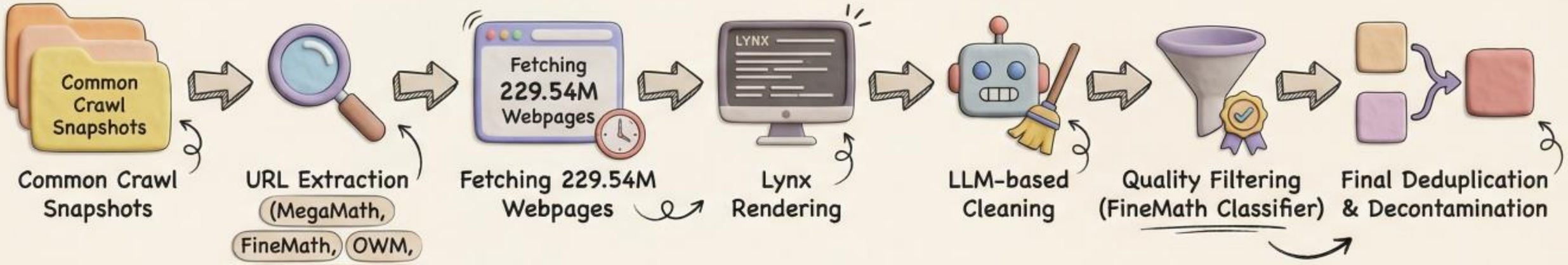
229.54M Webpages from 98 Common Crawl Snapshots (2014-2024)

Standardized, Cleaned, Preserved

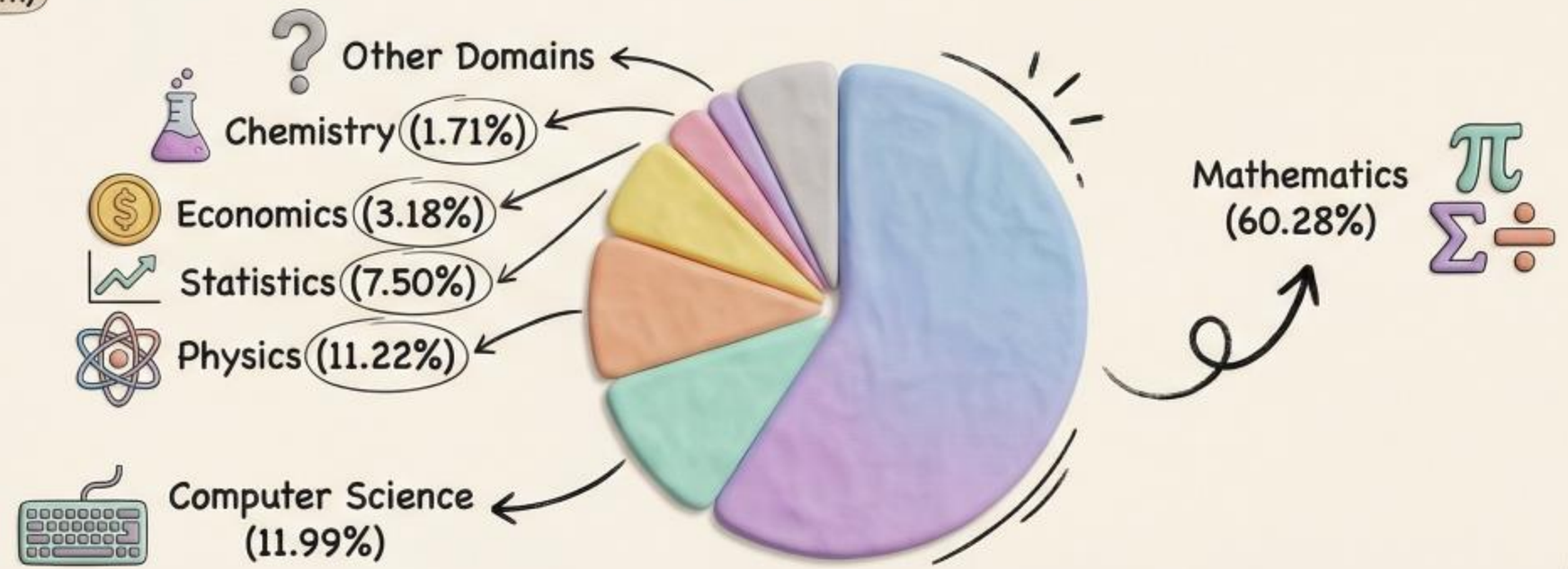
Key Pipeline

- Large-Scale Mathematical Corpus Construction
- Novel Domain-Agnostic Filtering
- High Quality from Common Crawl

The Extraction Pipeline: An Overview

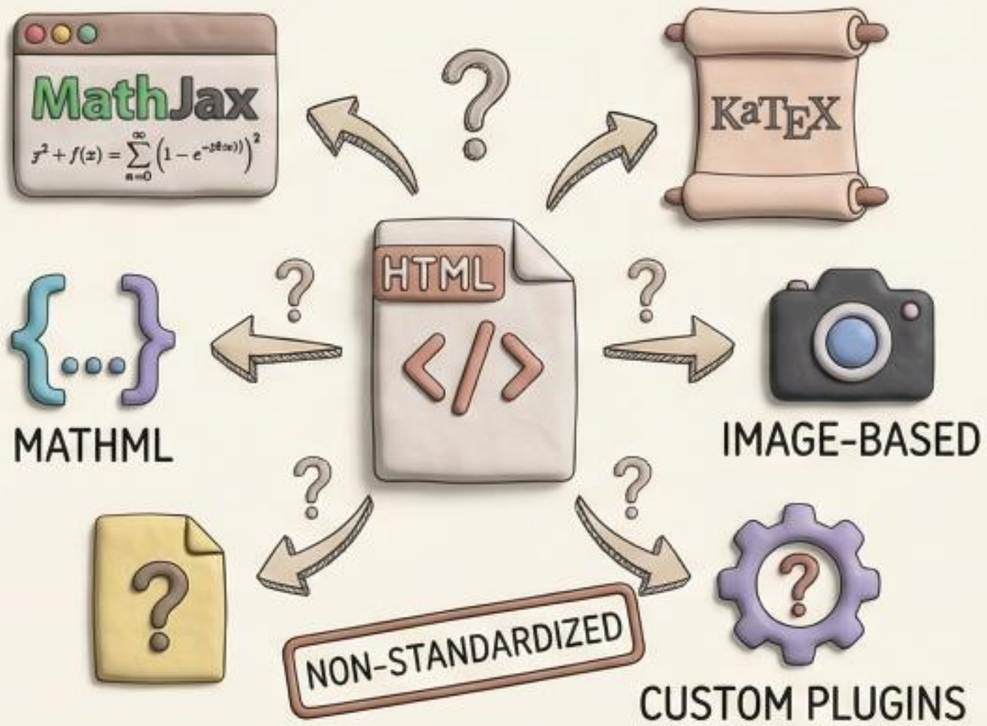


Topic Distribution



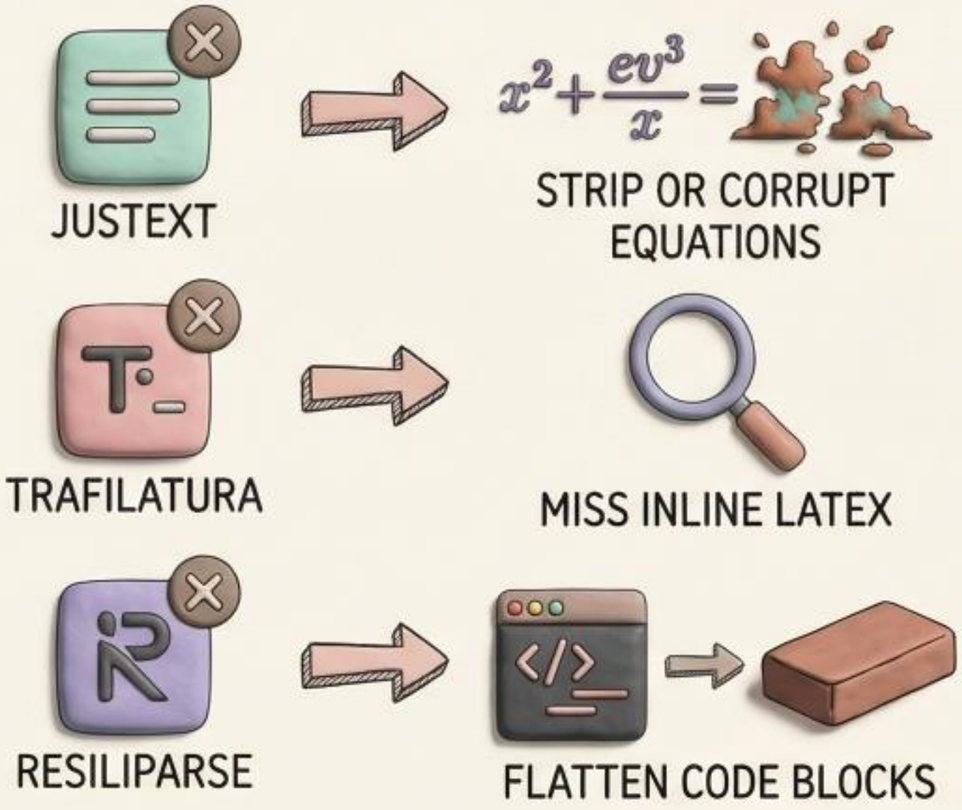
Reliable Text Extraction: Limitations of Prior Work

VARIABLE MATH REPRESENTATIONS



Highly variable forms across web; lack of conventions.

LIMITATIONS OF EXISTING TOOLS



Our Text Extraction Pipeline: The Two-Stage Approach

STAGE 1: Lynx Rendering & Preservation

Lynx Text-Based Browser

HTML → Plain Text

Output Mirroring Human Layout

STAGE 2: Phi-4 LLM Cleaning & Standardization

Phi-4 LLM (14B)

- Boilerplate Removed (Nav Bars, Headers)
- Standardized LaTeX Format
- Corrected Typographical Errors

MATHEMATICAL REPRESENTATION STANDARDIZATION

From Diverse HTML Formats to Unified LaTeX via LLM Conversion

Diverse HTML Input Formats

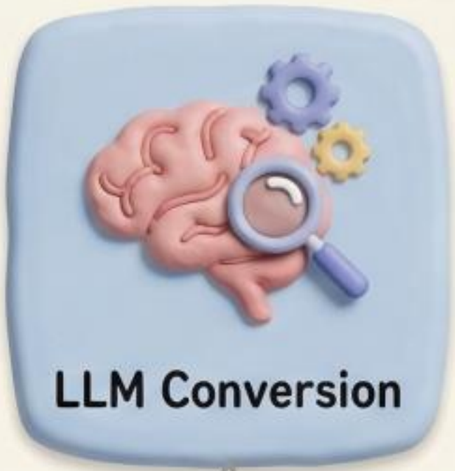
MathML (semantics & annotation)
<math xmlns="..."><semantics>...
</semantics></math>

pre tags (xml:lang='latex')
<pre xml:lang="latex">...</pre>

Image Tags (alt text LaTeX)
<img src="..." alt="...LaTeX...">

Inline LaTeX (custom delimiters)
\(... \) or [...]

Conversion Process



LLM Conversion

Ensures structural integrity & semantic consistency across the corpus

Standardized LaTeX Output

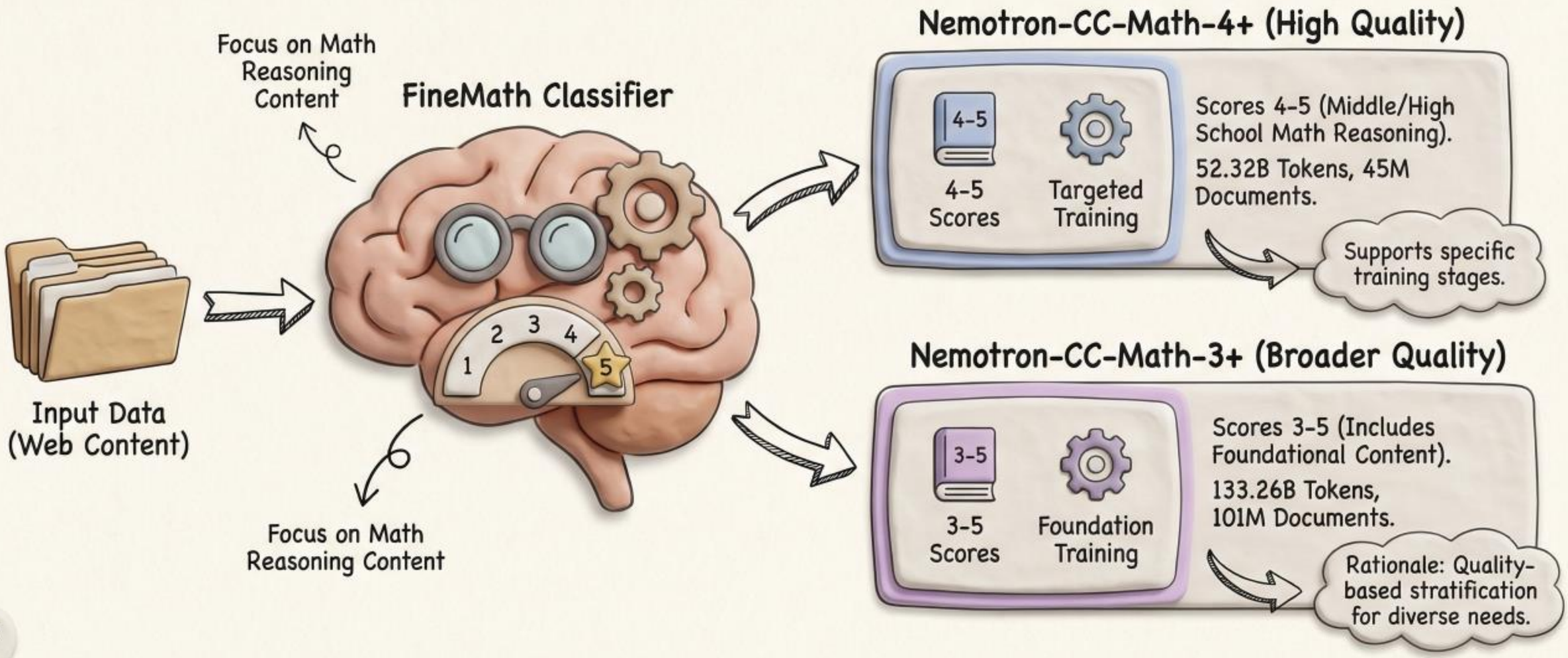
Unified LaTeX Representation

$$\frac{-b\,[-b \pm \sqrt{b^2 - 4ac}}{2a}$$

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

QUALITY CLASSIFICATION AND FILTERING

The FineMath Classifier & Dataset Stratification for Mathematical Reasoning



FUZZY DEDUPLICATION PROCESS

A Summary of Efficient & Secure Data Cleaning

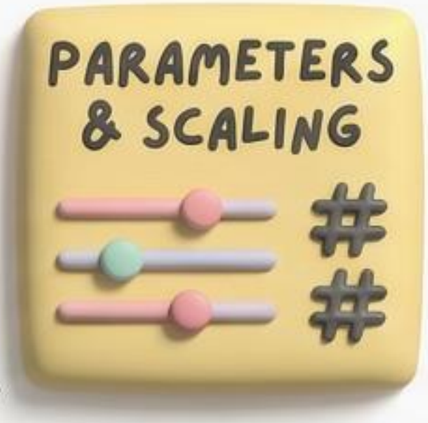
IMPORTANCE: EFFICIENCY & SECURITY

- Removes Near-Duplicates for Faster Training
- Reduces Memorization Risk

METHOD: MINHASH & LSH

24-Char N-Grams

Jaccard Threshold: 0.8



PARAMETERS & SCALING

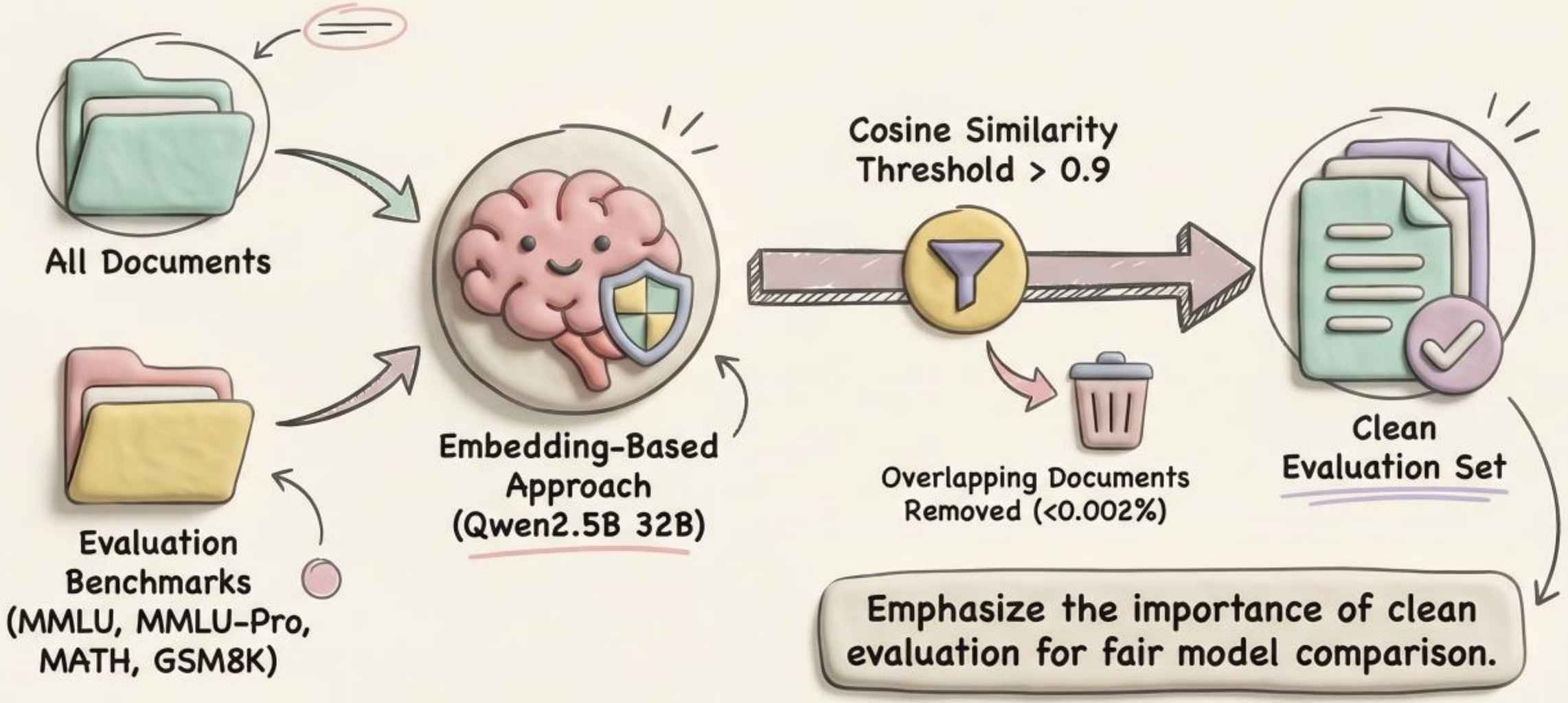
- r=20 bands
- b=13 functions/band
- Concurrent Shuffling (5 bands)

RESULT: WEB-SCALE DEDUPLICATION

Efficient Identification

Handles Massive Datasets

DECONTAMINATION AGAINST BENCHMARKS



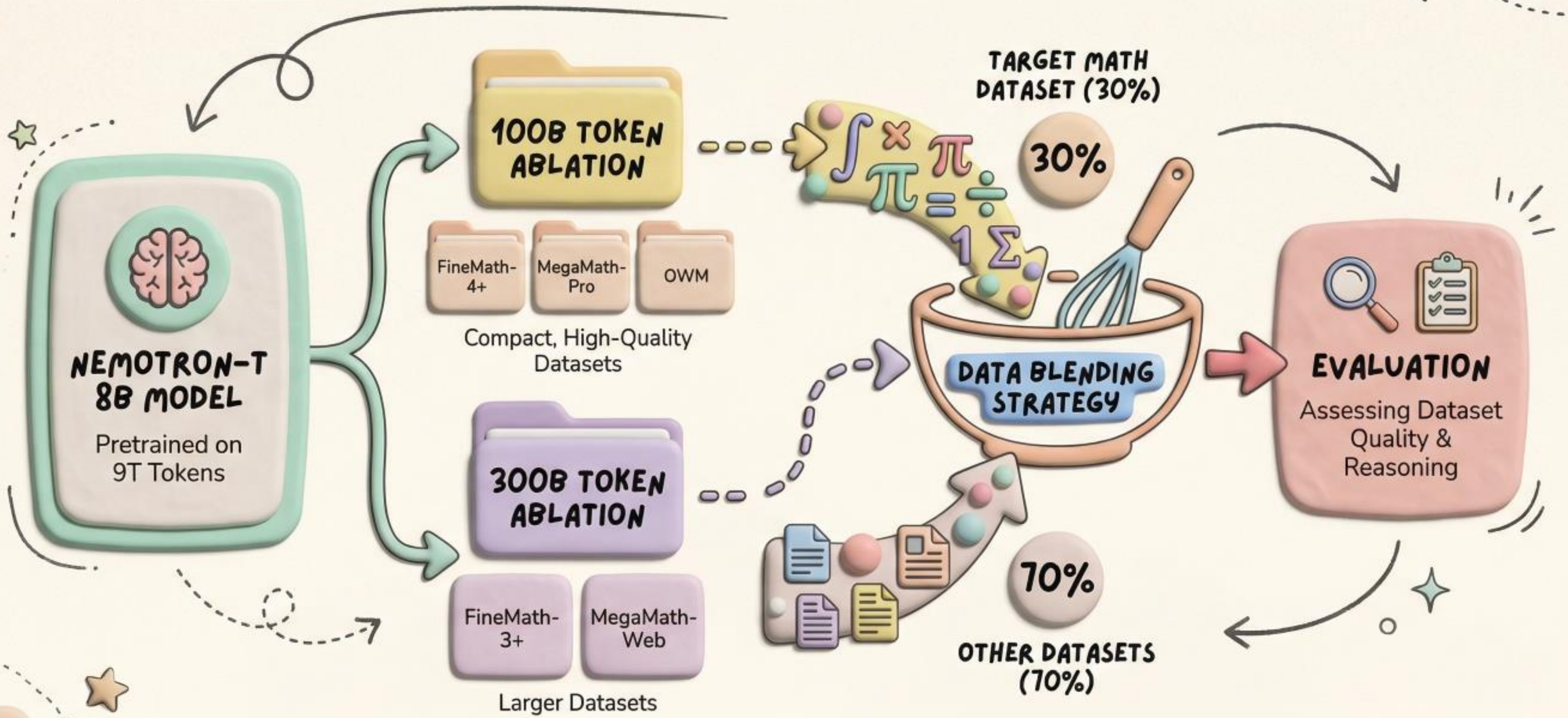
DATASET COMPARISON: SCALE AND QUALITY

Dataset	Document Count	Token Count	Source Types	License & Accessibility
Nemotron-CC-Math-4+	45.1M Docs	52.32B Tokens	Common Crawl, Web, Math Repos	Permissive (Common Crawl)
FineMath-4+	Various Docs	9.5B Tokens	Common Crawl, Web	Various / Limited
MegaMath-Pro	Various Docs	14.7B Tokens	Common Crawl, Web, Math Repos	Various / Limited

5.5x Larger Scale than previous best (FineMath-4+)!

Broad Accessibility & Permissive License

EXPERIMENTAL SETUP: ANNEALING ABLATIONS



BENCHMARKS AND EVALUATION METRICS

KNOWLEDGE UNDERSTANDING

MMLU-Pro
MMLU
MMLU-STEM

Exact Match Accuracy

CODE GENERATION

MBPP
HumanEval
MBPP+
HumanEval+

avg@20 metric
(Nucleus Sampling, Temp 0.6, Top-p 0.95)

MATHEMATICAL REASONING

GSM8K
MATH

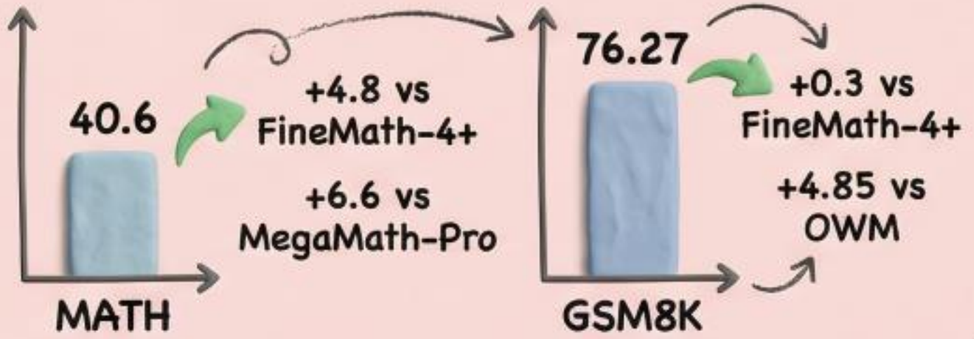
Greedy Decoding with Math-Verify
(Symbolic Matching)

RIGOROUS EVALUATION APPROACH

- Ensures fair comparison across domains
- Standardized procedures for all models
- Transparent methodology for reproducibility

100B Token Results: Comprehensive Performance Analysis

MATH TASKS



CODE PERFORMANCE

34.82

HumanEval+
(leads most benchmarks)

+2.3 vs OWM

KNOWLEDGE TASKS

38.49

MMLU-Pro

+2.1 vs MegaMath-Pro

Consistent Dominance

CROSS-DOMAIN BOOST

High-Quality Math Data Boosts Performance Across Domains

300B Token Results: Scaling Performance Gains

MATH BENCHMARK

Nemotron-CC-Math-3+ vs MegaMath-Web

44.2

+9.6 vs FineMath-3+
+12.6 vs MegaMath-Web

GSM8K 80.06 (+0.6 vs FineMath-3+, +3.6 vs OWM)

CODE GENERATION

MBPP+

43.51

+4.6 vs MegaMath-Web,
+14.32 vs FineMath-3+

HumanEval+ 37.16 (+3.0 vs FineMath-3+)

GENERAL KNOWLEDGE

MMLU-STEM

64.26

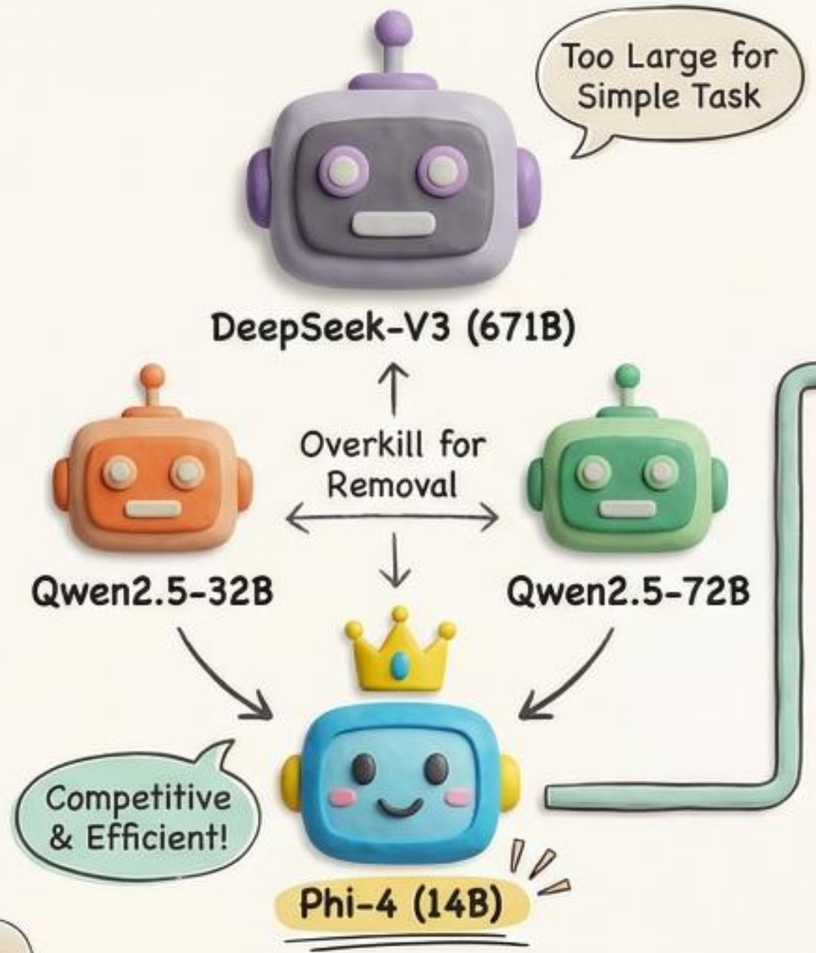
Showing cross-domain transfer capabilities

SCALING BENEFITS PROVEN

Gains scale with increased pretraining on high-quality data.

Model Choice Ablation: Efficiency vs Performance

Comparing LLMs for **Webpage Boilerplate Removal** across 7M Documents



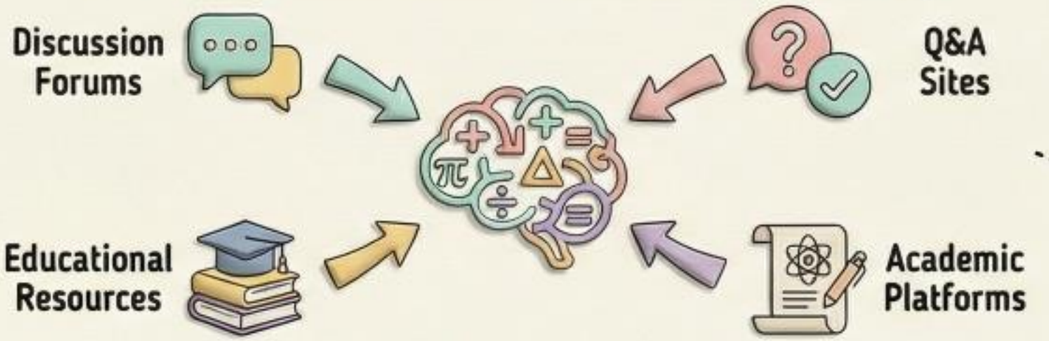
Model	Math (GSM8K)	Math (MATH)	Computational Cost
Phi-4 (14B)	✓ 79.98	✓ 40.6	Substantially Lower
DeepSeek-V3 (671B)	Competitive	Competitive	High
Qwen2.5-32B	Competitive	Competitive	High
Qwen2.5-72B	Competitive	Competitive	High

Conclusion:

Webpage boilerplate removal doesn't require excessively large models, enabling efficient large-scale processing.

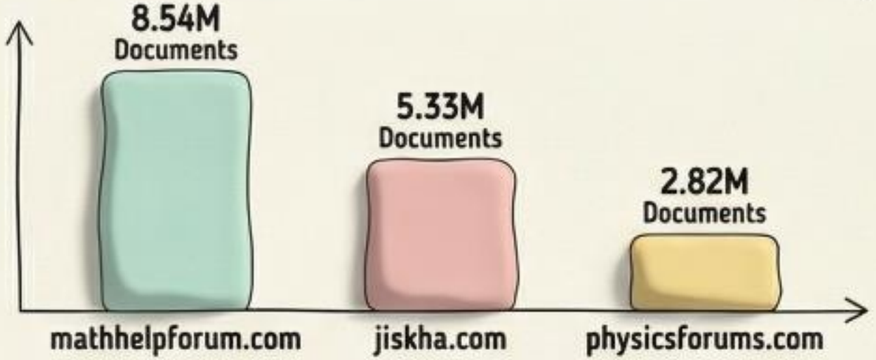
DATASET COMPOSITION ANALYSIS

DATASET COMPOSITION OVERVIEW



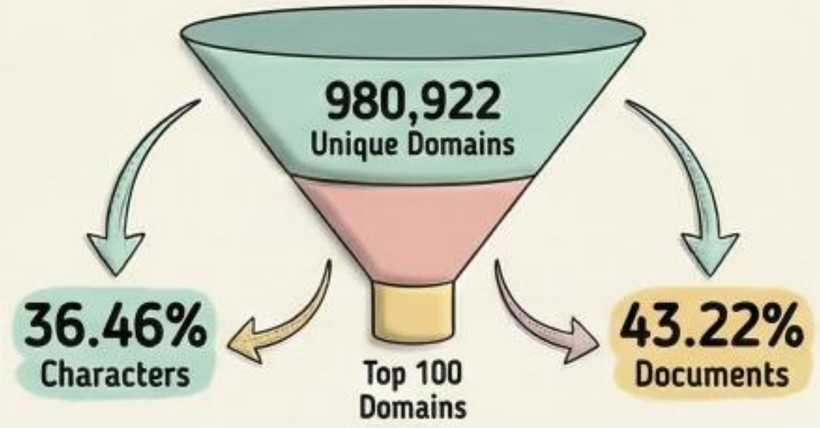
01 Diverse sources including forums, Q&A, educational, and academic sites.

TOP 20 DOMAINS (DOCUMENT COUNT)



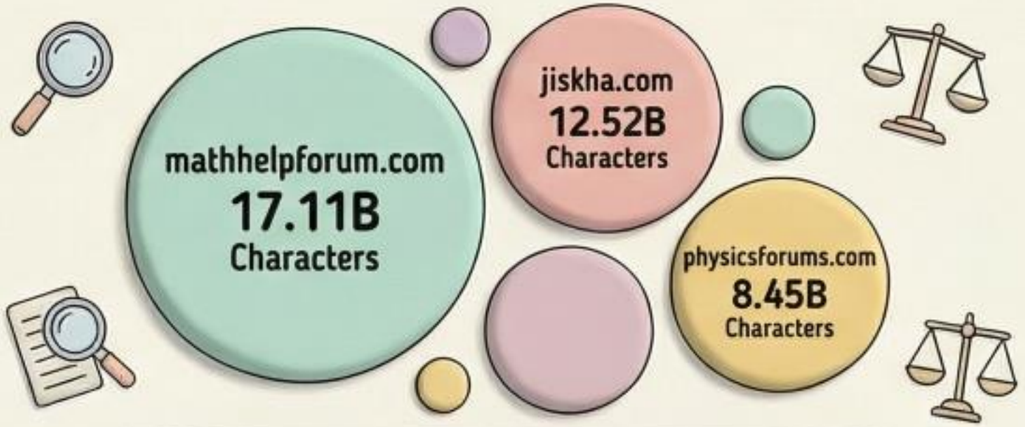
02 Showing the top three domains by number of documents.

DOMAIN SPAN & CONCENTRATION



05 Significant concentration in top domains despite large overall span.

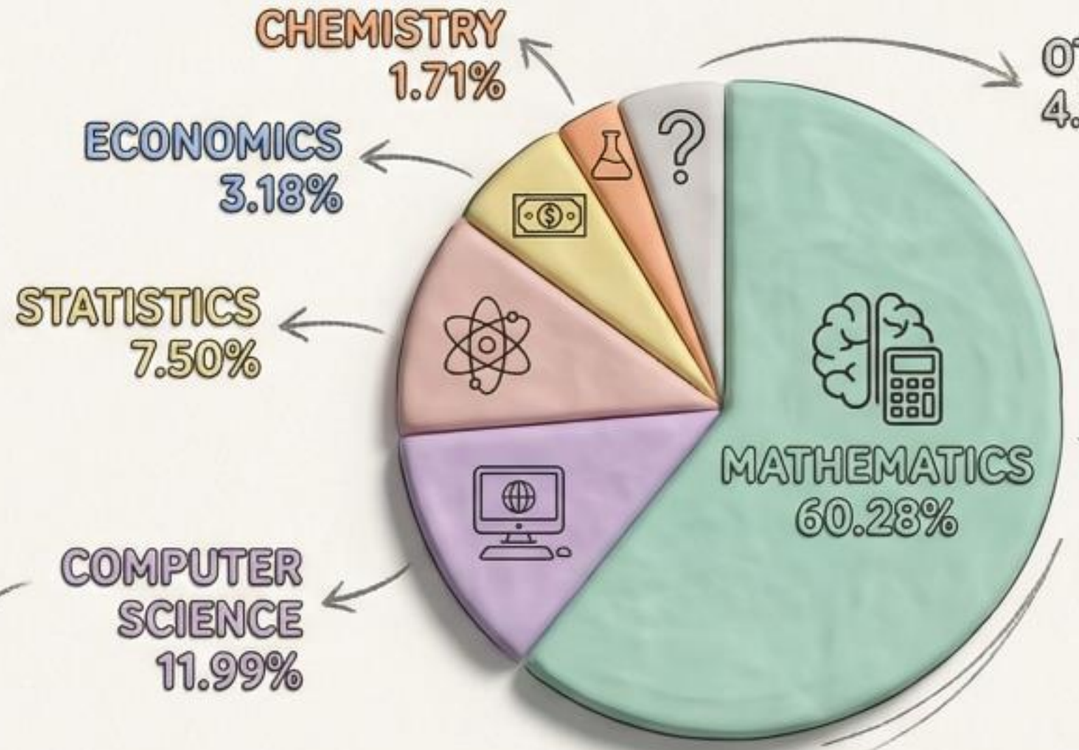
TOP 20 DOMAINS (CHARACTER COUNT)



04 Highlighting the top three domains by total character count.

TOPIC DISTRIBUTION & CONTENT DIVERSITY

Analysis of 150,000 randomly sampled documents via Qwen3-30B-A3B-Instruct-2507.



DIVERSE SUBJECT COVERAGE

Enables broad scientific reasoning capabilities beyond pure mathematics.

BROAD SCIENTIFIC REASONING

Supports interdisciplinary knowledge transfer and connections.

INTERDISCIPLINARY KNOWLEDGE

Fosters holistic understanding and problem-solving across fields.

Qualitative Analysis: Degenerate Cases in Prior Work

High Scores, Low Quality: The Paradox of MegaMath-Pro

Repetitive, low-quality content achieves unexpectedly high mathematical & language scores.

Degenerate Examples ✗

The angle will be calculated and displayed...
The Integral Calculator is able to calculate integrals online...
The angle will be ca[illegible]d and displayed...
The Integral Calcul[illegible] and displayed...
The Integral Calculator is able to calculate integrals online...

Artifacts raise concerns about dataset reliability & confirm the importance of robust quality filtering.

Robust Quality Filtering

SIDE-BY-SIDE COMPARISON: CODE PRESERVATION

Nemotron-CC-Math (Superior Preservation)

```python
class SparseMatrix:
    def __init__(self):
        self.data = {}

    def set(self, r, c, v):
        self.data[(r, c)] = v

    def mult(self, other):
        result = SparseMatrix()
        for (r, c), v in self.data.items():
            result.set(...) # Code with Proper Indentation
```

Perfect Indentation & Structure Preserved. Critical for Python.

Incidental Code Data
4.3M (3+) | 1.44M (4+)

Boosts Generation Performance

Comparison

Prior Work (Stripped/Corrupted)

```python
class SparseMatrix:
  def __init__(self):
    self.data = {}

  def set(self, r, c, v):
    self.data[(r, c)] = v

  def mult(self, other):
    result = SparseMatrix()
    for (r, c), v in self.data.items():
      result.set(....)
```

Stripped Formatting, Lost Indentation. Corrupted Structure.

Limited/ Stripped Data

Poor Generation

Impact on Reasoning Performance: Superior Code Generation due to Preserved Structure

SIDE-BY-SIDE COMPARISON: MATHEMATICAL EQUATIONS

Nemotron-CC-Math: High-Quality Preservation

Inline Equation

$E = mc^2$ → Properly Rendered

Display Equation

$\int f(x)dx = F(x) + C$ → Clear Semantic Meaning

Matrix Representation

$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ → Essential for Reasoning

Maintaining semantic meaning essential for mathematical reasoning.

Preservation Pipeline vs. Corruption

Prior Work (OpenWebMath, MegaMath): Lost/Corrupted Notation

Inline Equation

$E = m?2$ → Corrupted Input

Display Equation

$\int_x^2 f(x)dx = ? + C$ → Lost Meaning

Matrix Representation

$\begin{bmatrix} a & ? \\ c & \end{bmatrix}$ → Reasoning Inhibited

Often removes or corrupts mathematical expressions.

RELATED WORK: OPEN-SOURCE DATASETS

OpenWebMath: 14.7B Tokens

Limitation: Brittle heuristics, rendering often corrupts formulas.

InfiMM-WebMath: 40B Tokens

Multimodal dataset

FineMath: 54B Tokens

Limitation: Inherits OWM pipeline issues

Proof-Pile & Proof-Pile-2: 8.3B & 55B Tokens

MathPile: 9.5B Tokens

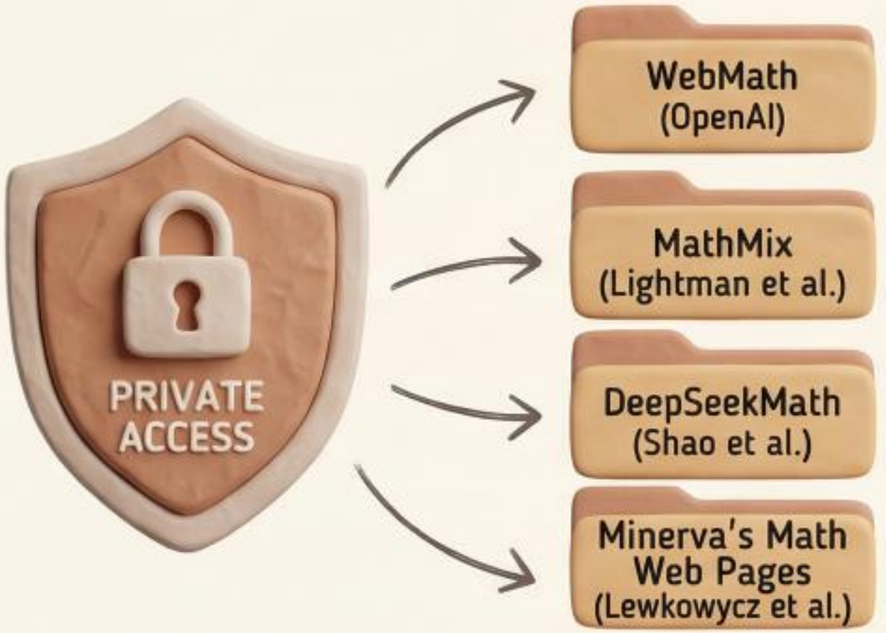
arXiv
Textbooks
Forums

Limitation: Much content remains raw LaTeX.

NEMOTRON-CC-MATH SOLUTION

Addressing limitations through improved pipeline, quality filters, and scale.

RELATED WORK: PROPRIETARY DATASETS



Demonstrate impressive capabilities but lack public access, limiting reproducibility.

VS.

Fosters community progress and enables reproducible research in mathematical reasoning.

Key Contributions and Innovations

1. Reliable Extraction Pipeline

First to reliably extract scientific content & math from noisy web-scale data.

2. Largest Open Math Corpus

133B tokens, 5.5x larger than previous best, 4+ subset.

3. Full Pipeline Open-Sourced

For reproducibility and domain adaptation.

4. Comprehensive Analysis

Composition, sources, and topic distribution.

5. Superior Performance

Across math, code, and knowledge benchmarks.

IMPACT ON MATHEMATICAL REASONING PERFORMANCE

MATH BENCHMARK

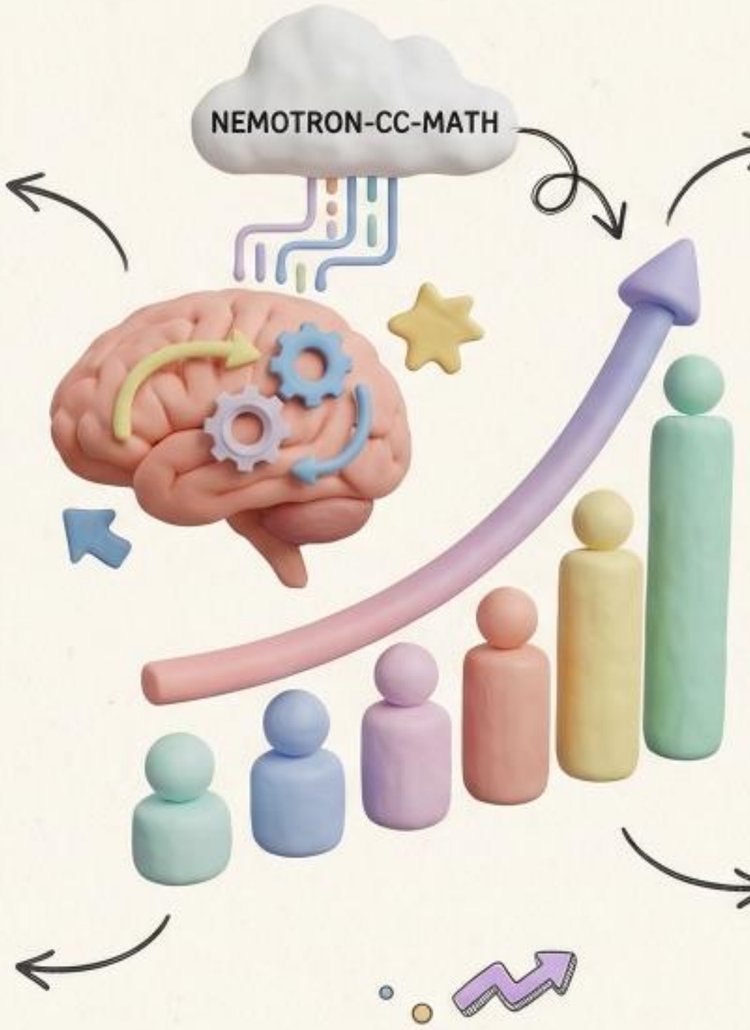
+4.8 to +12.6 GAIN

Gains on MATH benchmark over strong baselines. Significant improvement in mathematical problem-solving capabilities.

GENERAL KNOWLEDGE

Improvements on MMLU & MMLU-STEM

Showing cross-domain transfer. Enhances performance across broader subjects.



CODE GENERATION

+4.6 to +14.3 GAIN

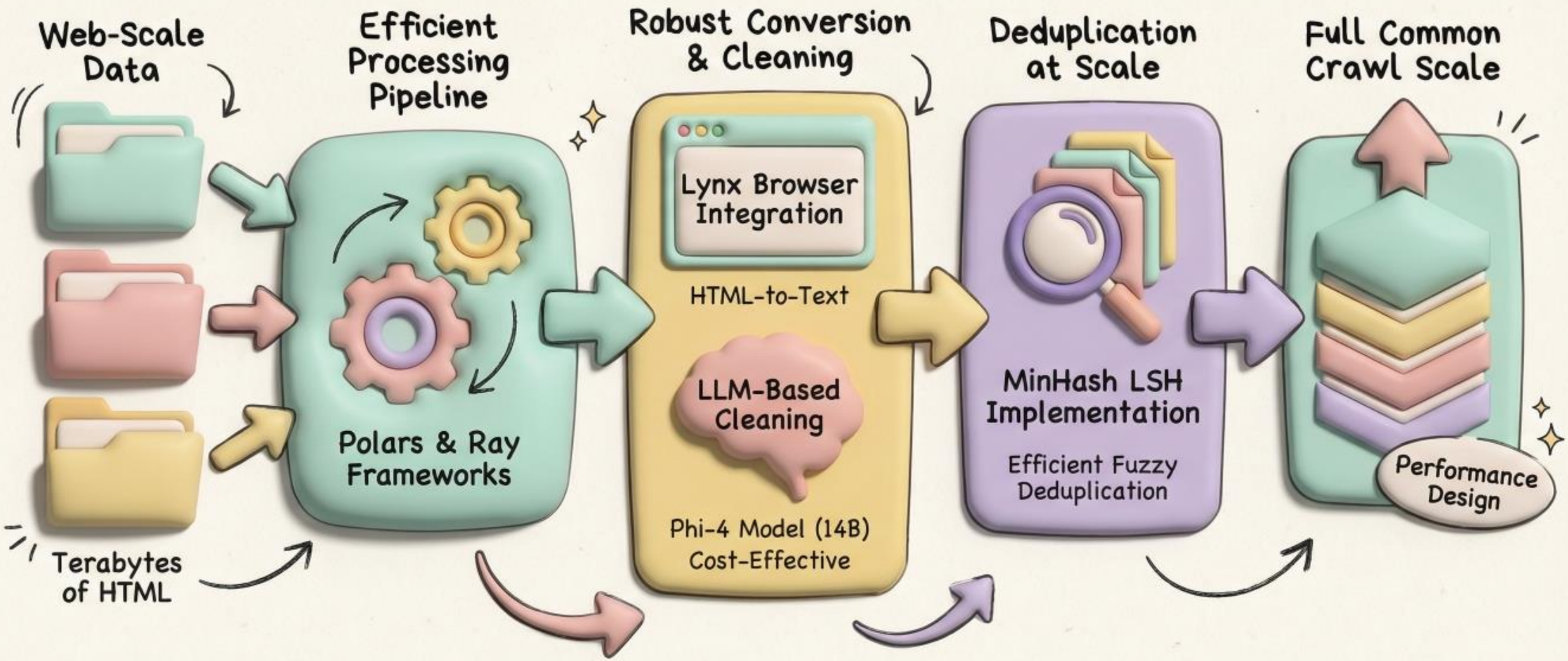
Gains on MBPP+ benchmark despite not explicitly targeting code. Shows transfer to algorithmic tasks.

BROADER REASONING

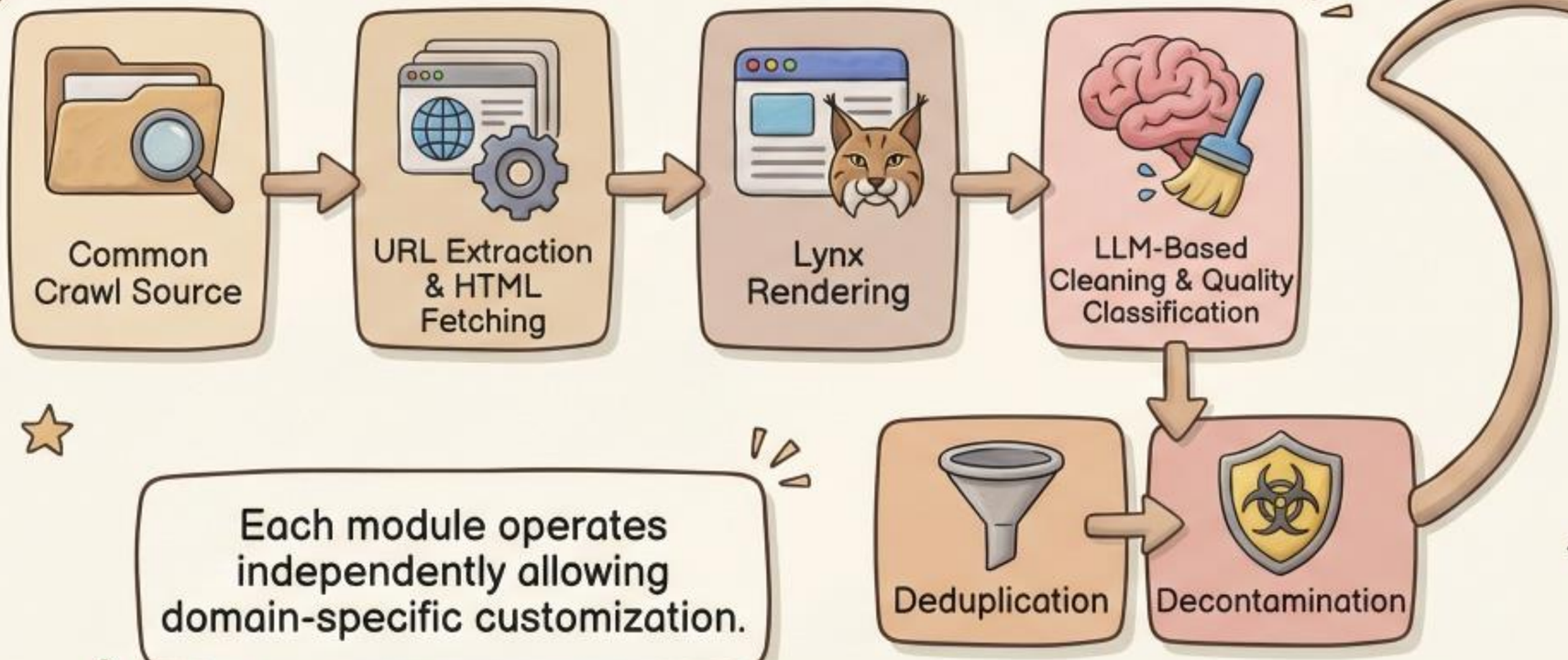
STRENGTHENS SKILLS

High-quality mathematical data strengthens broader reasoning skills beyond domain-specific tasks.

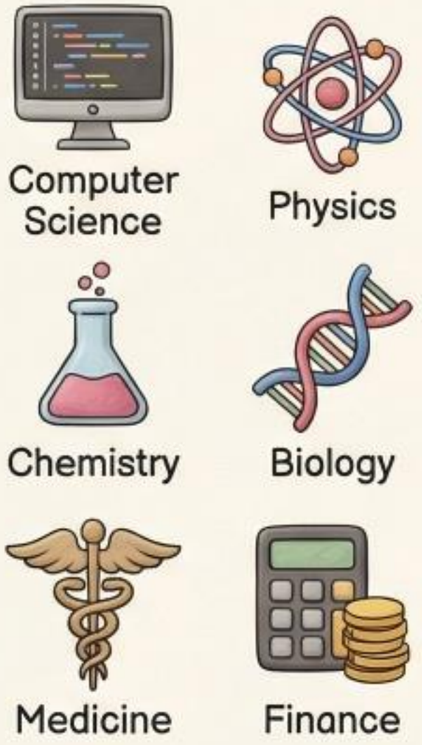
Scalability and Performance Optimization



DOMAIN-AGNOSTIC DESIGN AND FUTURE APPLICATIONS



Each module operates independently allowing domain-specific customization.



Demonstrates how robust scientific text extraction can benefit diverse technical fields.

LIMITATIONS AND FUTURE WORK

CURRENT LIMITATIONS

<1% Common Crawl (URL-based)

Inherited Quality Classifier (Limited Math Types)

English-Dominant (Multilingual Reasoning Gap)

OPPORTUNITIES FOR IMPROVEMENT

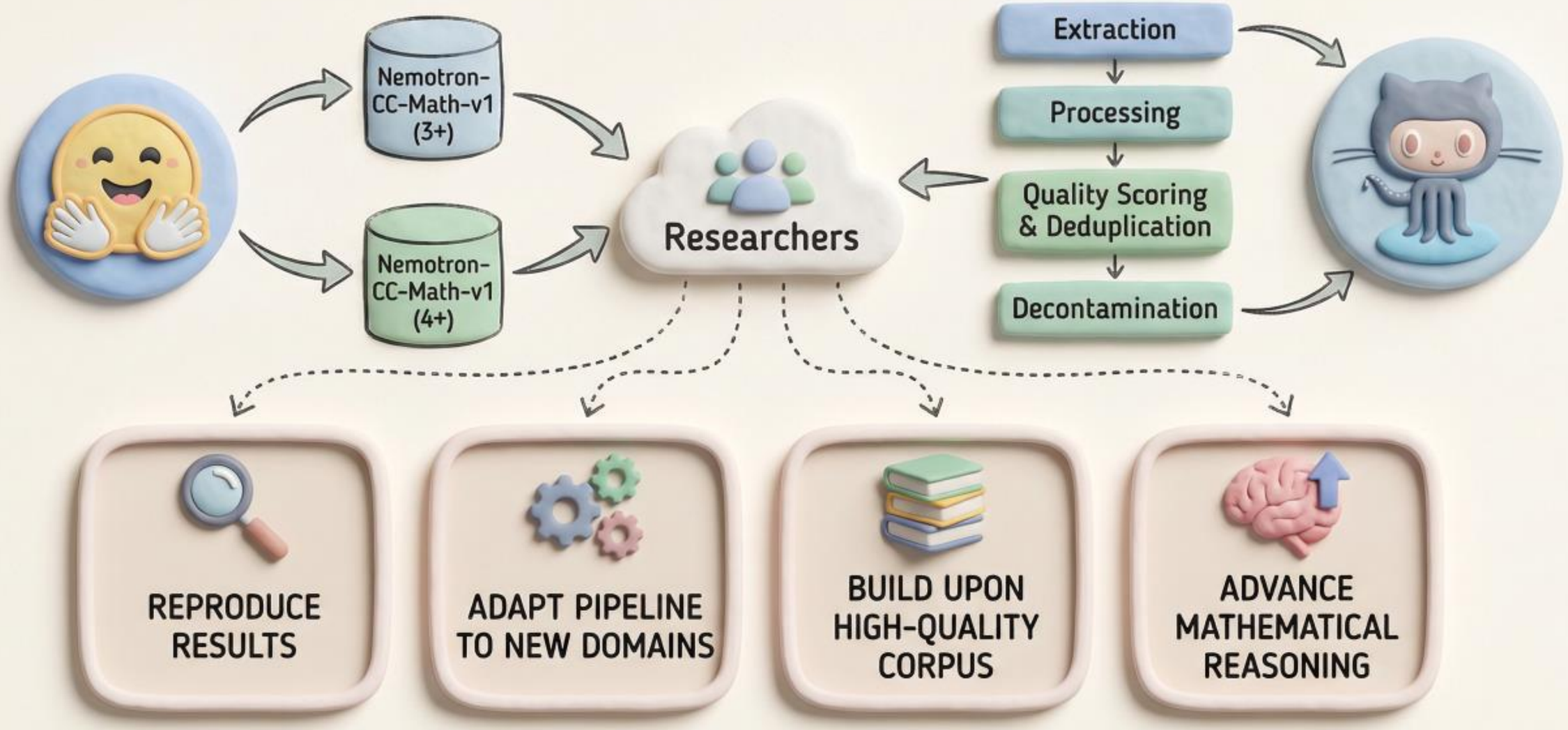
Multimodal Math Data Integration

Domain-Specific Quality Classifiers

Expansion to Scientific Domains

Integration with Formal Theorem Proving Systems

OPEN SOURCE RELEASE AND COMMUNITY IMPACT



Conclusion: Advancing Mathematical Reasoning at Scale

Highest Quality Open-Source Math Corpus

Nemotron-CC-Math: Best-in-class foundation, enabling measurable gains.

Enabling Measurable Gains Across Domains

Mathematical Reasoning

Code Generation

General Knowledge (STEM)

Novel Extraction Pipeline Addresses Challenges

Extraction Pipeline

Future Implications & Open Strategy

- Crucial Foundation for Specialized Reasoning Models
- Promotes Transparency & Community Collaboration