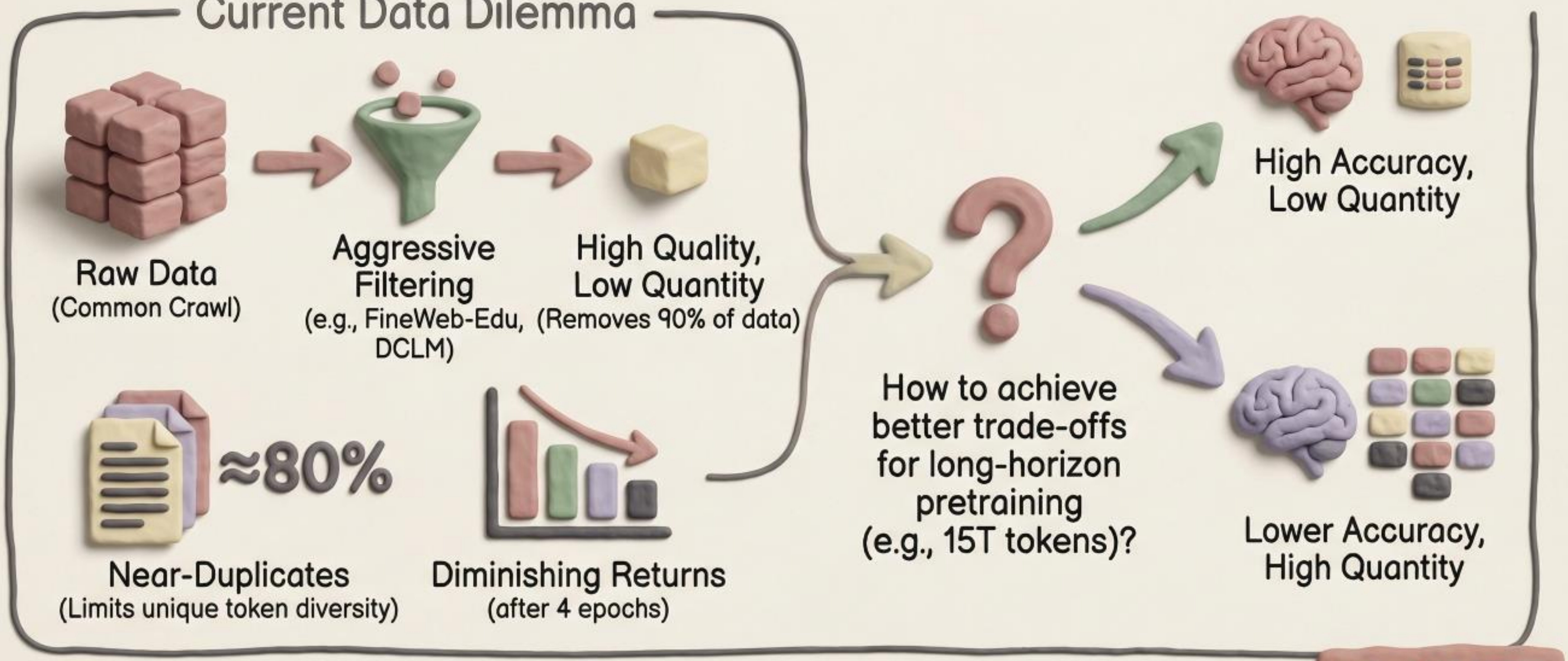

Nemotron-CC:
Transforming Common Crawl
into a Refined Long-Horizon
Pretraining Dataset

The Challenge: Long-Horizon LLM Training



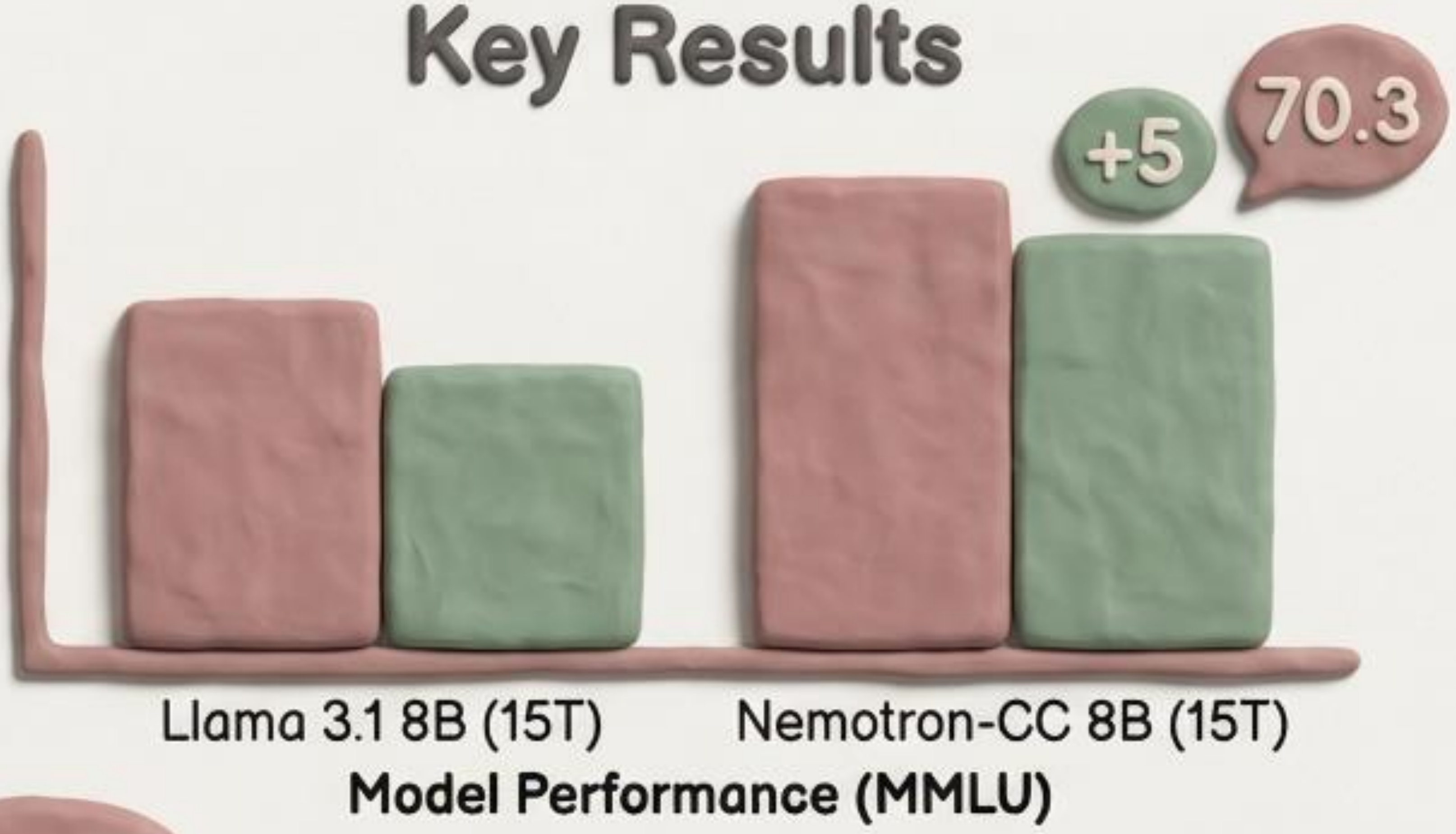
arXiv:2412.02595

Research Contributions & Key Results

Main Contributions

1. 6.3T token dataset transformation method
2. Proven effectiveness through comprehensive comparisons
3. Detailed ablation studies revealing best practices

Key Results



15T Tokens

1.1T high-quality subset: **+5.6** MMLU over DCLM

Full **6.3T** dataset matches DCLM, **4x** more unique tokens

Nemotron-CC 8B (15T)

+3.1 ARC-Challenge

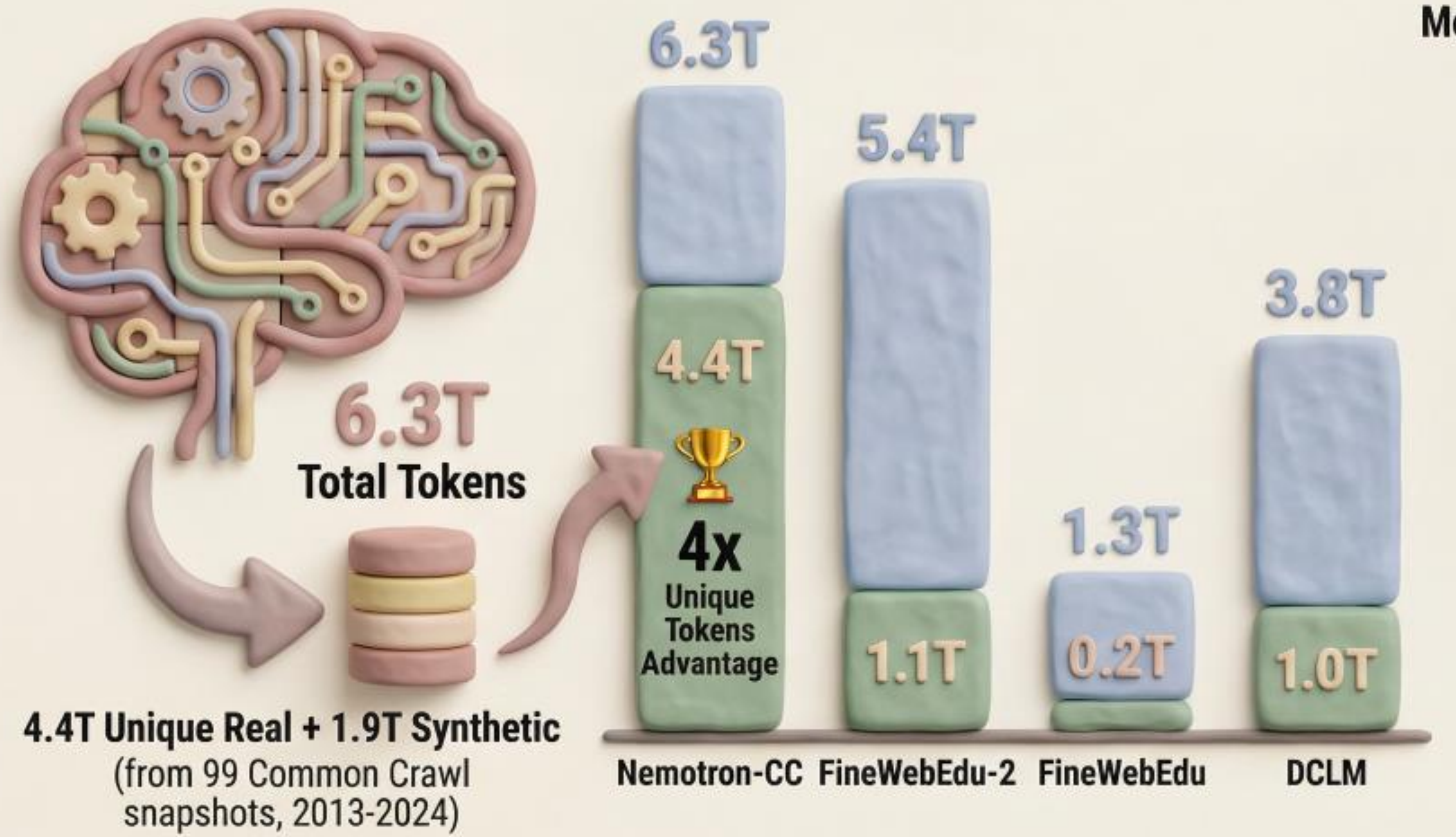
+0.5 Average (10 tasks)

Outperforms Llama 3.1 8B

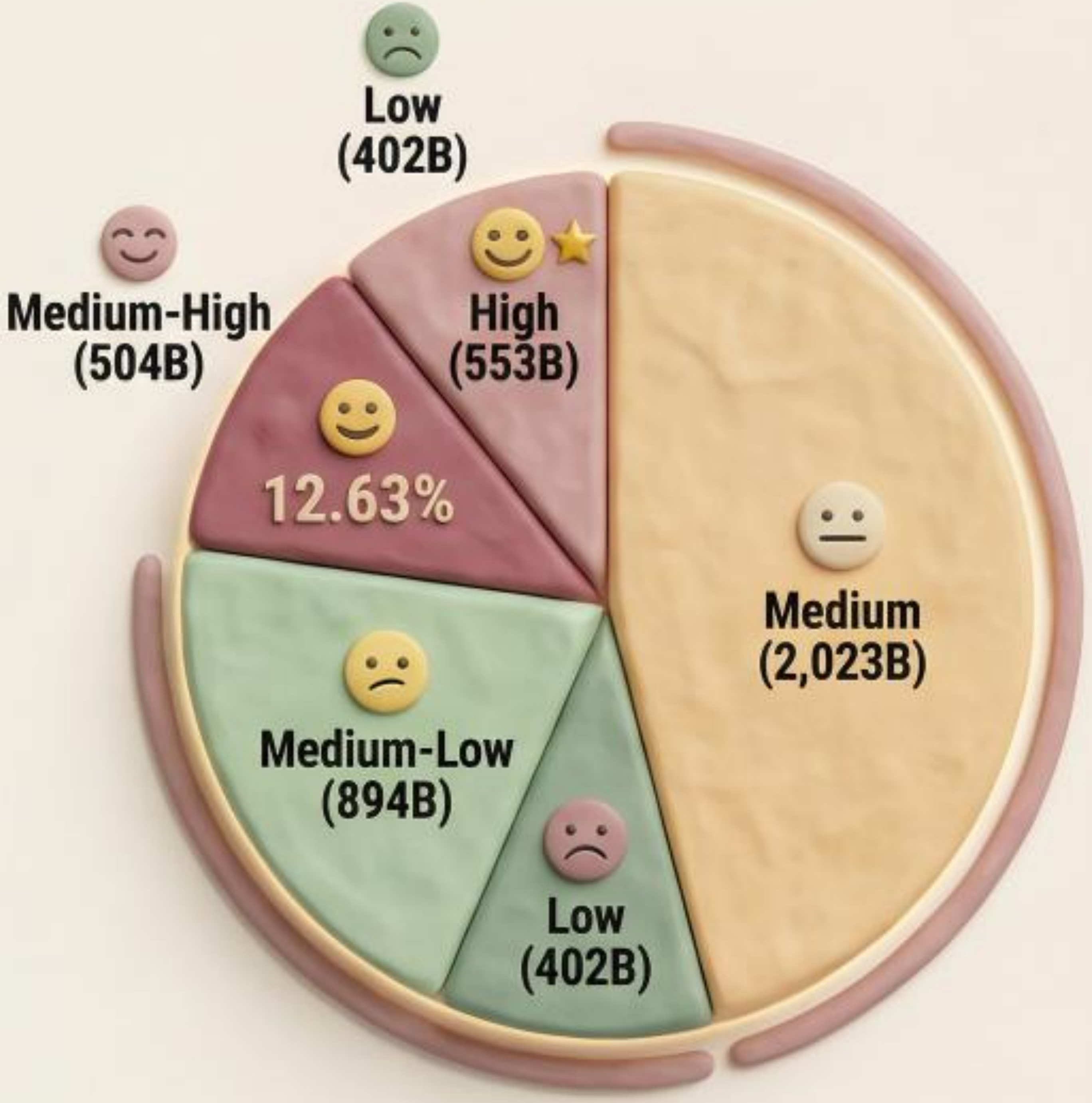
Dataset Overview: Scale & Composition

Unprecedented Scale with a Focus on Unique, High-Quality Data

Total Size & Unique Tokens

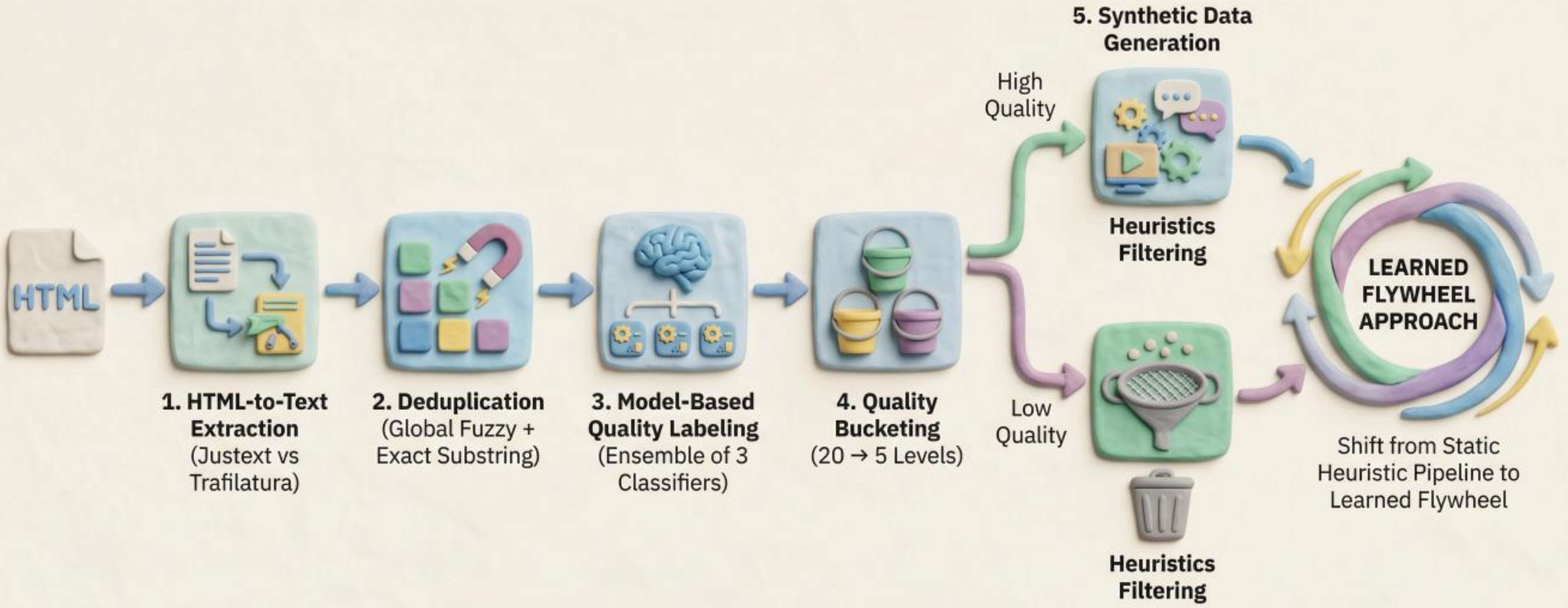


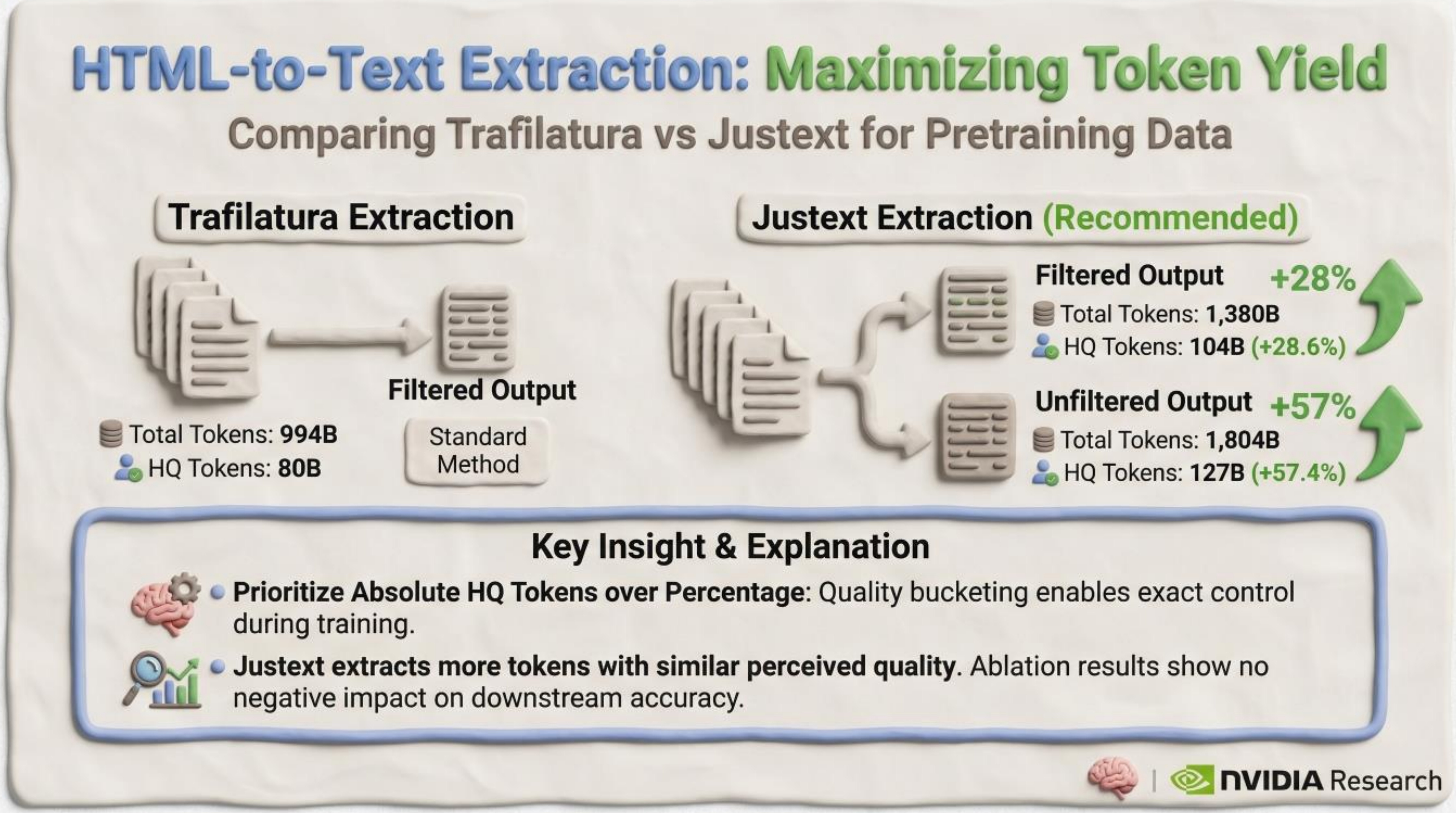
Quality Distribution



Over 70% of the dataset is Medium quality or higher, ensuring robust training.

The Nemotron-CC Pipeline: An Overview

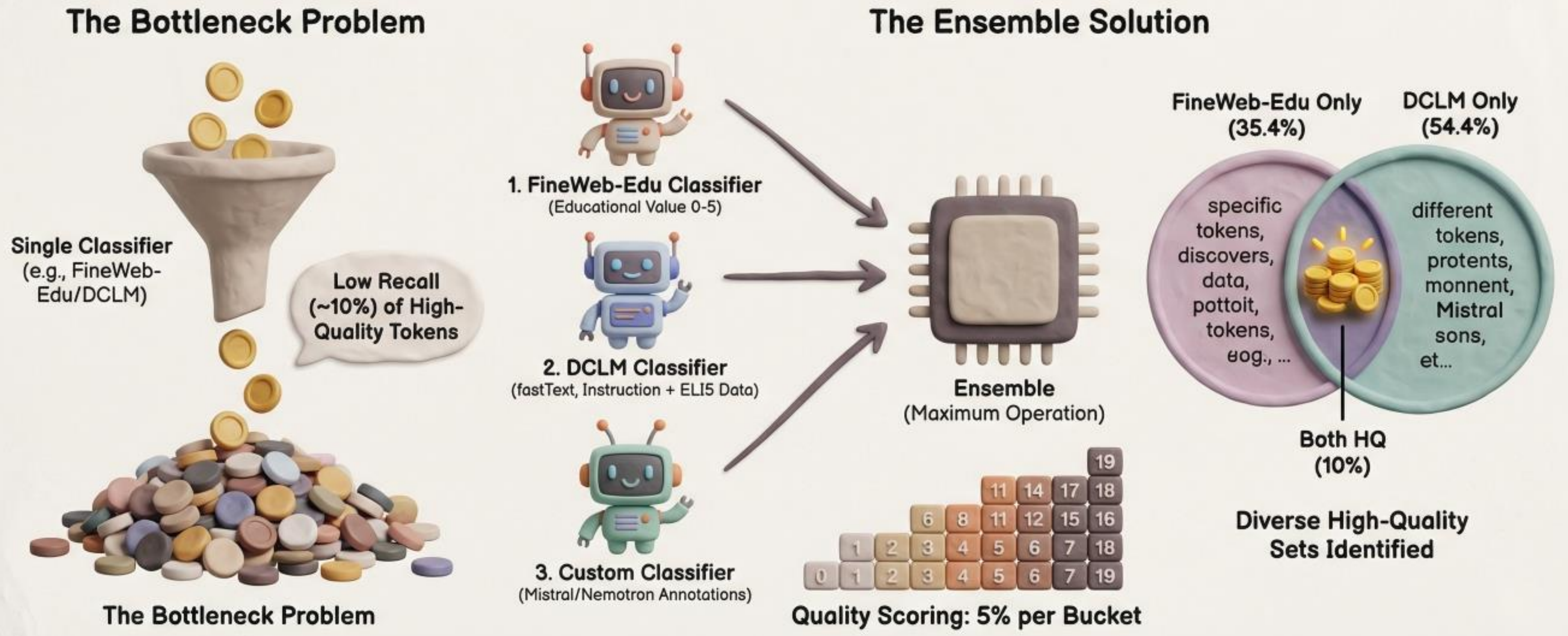


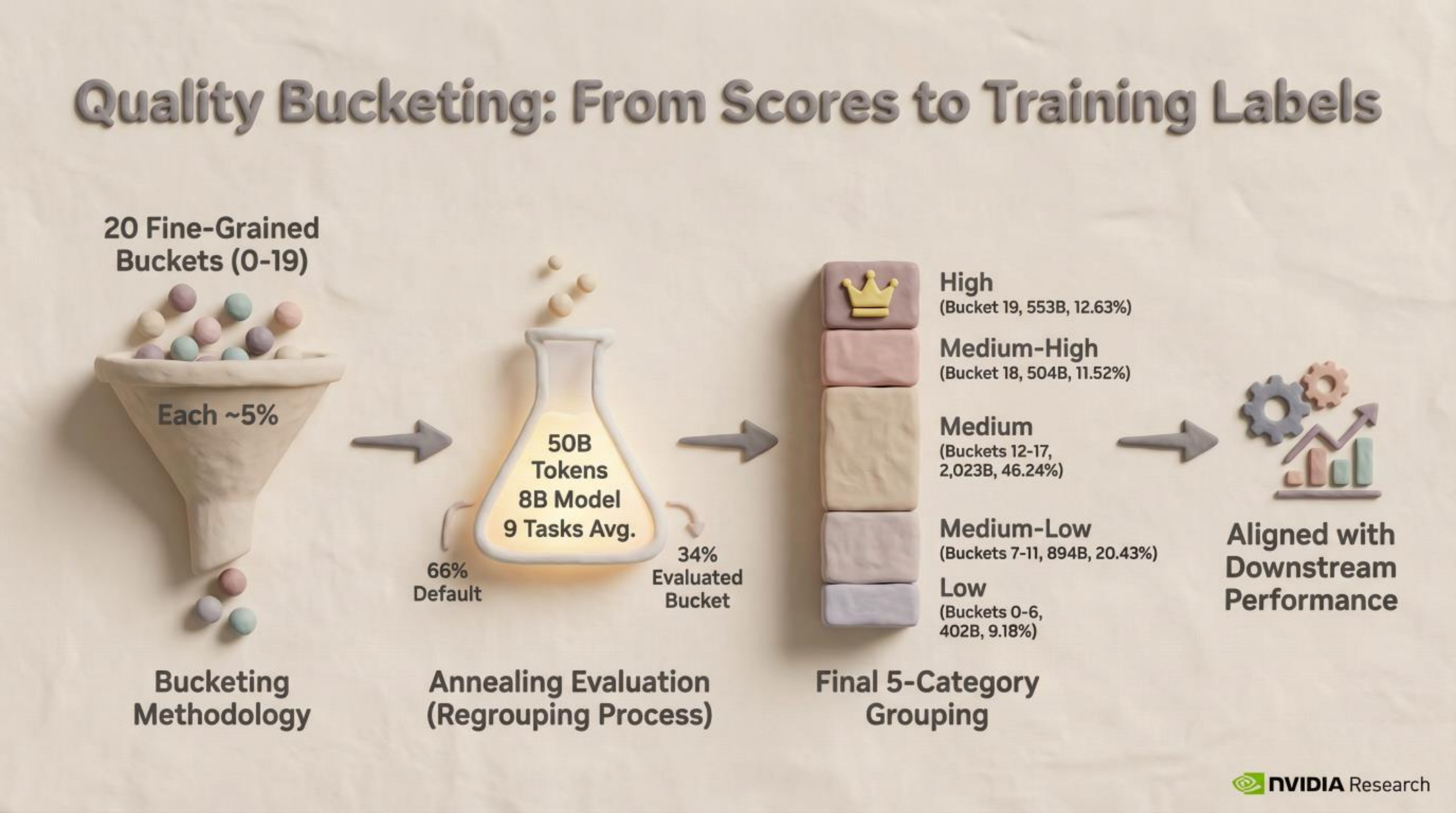

HTML-to-Text Extraction: Maximizing Token Yield
Comparing Trafilatura vs Justext for Pretraining Data
Trafilatura Extraction
Filtered Output
Total Tokens: 994B
HQ Tokens: 80B
Standard Method
Justext Extraction (Recommended)
Filtered Output
+28%
Total Tokens: 1,380B
HQ Tokens: 104B (+28.6%)
Unfiltered Output
+57%
Total Tokens: 1,804B
HQ Tokens: 127B (+57.4%)
Key Insight & Explanation
Prioritize Absolute HQ Tokens over Percentage: Quality bucketing enables exact control during training.
Justext extracts more tokens with similar perceived quality. Ablation results show no negative impact on downstream accuracy.
NVIDIA Research

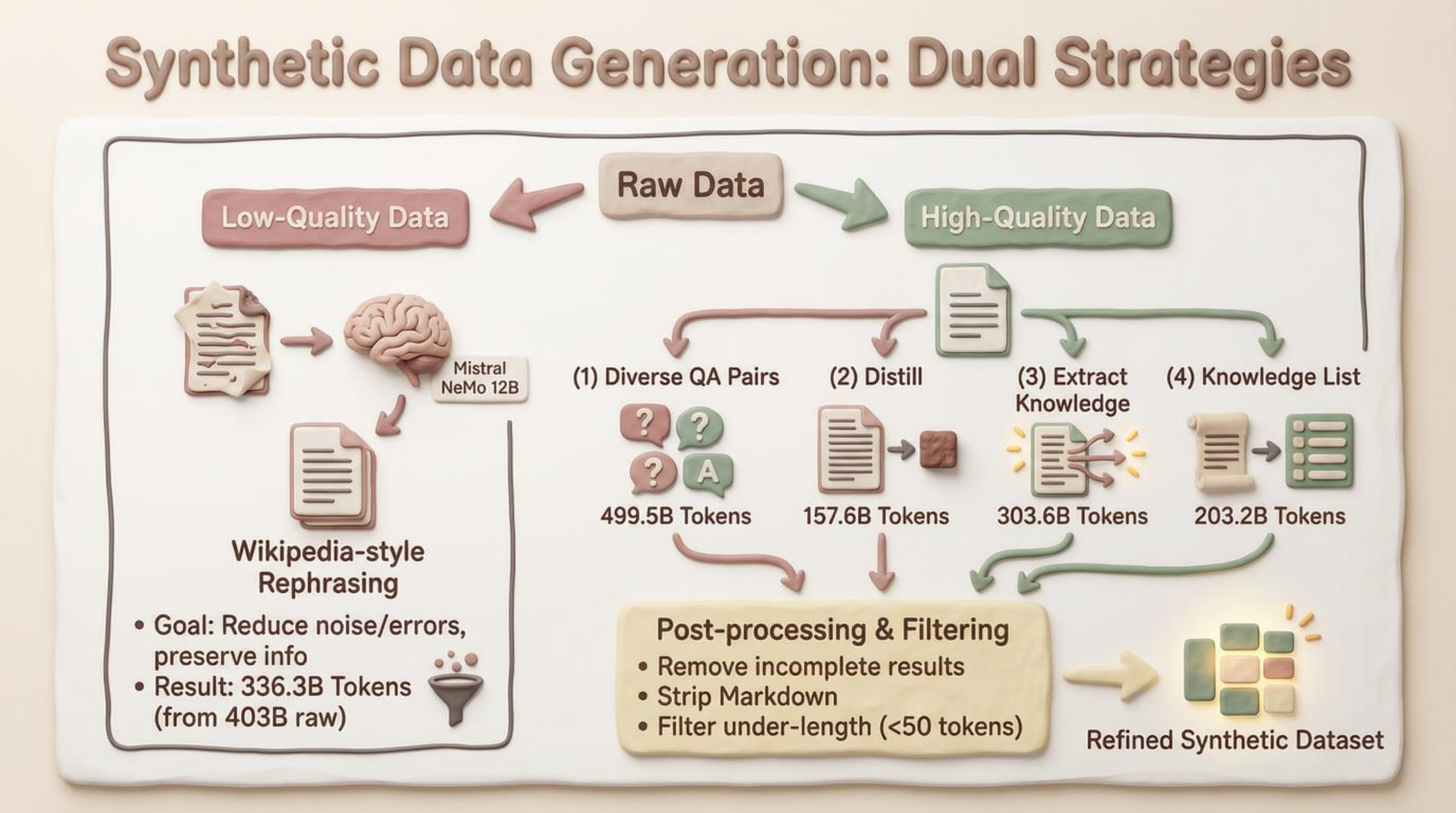

Heuristic Filtering: A Strategic Reevaluation
Conventional Approach
English Language Detection
Perplexity Filter (KenLM)
Document Length Constraints
Symbol/Word Ratio Filters
Critical Finding: Filtering removes 18.1% of High-Quality Tokens (FineWeb-Edu)
Removed Tokens (Low & High Quality)
Novel Strategy
Model-Based Classifiers (e.g., FineWeb-Edu)
High-Quality Tokens
Disable Heuristic Filters
Low-Quality Tokens
Apply Filters
Ablation Results: Justext HQ-unfiltered Improves MMLU by +2% vs. Filtered

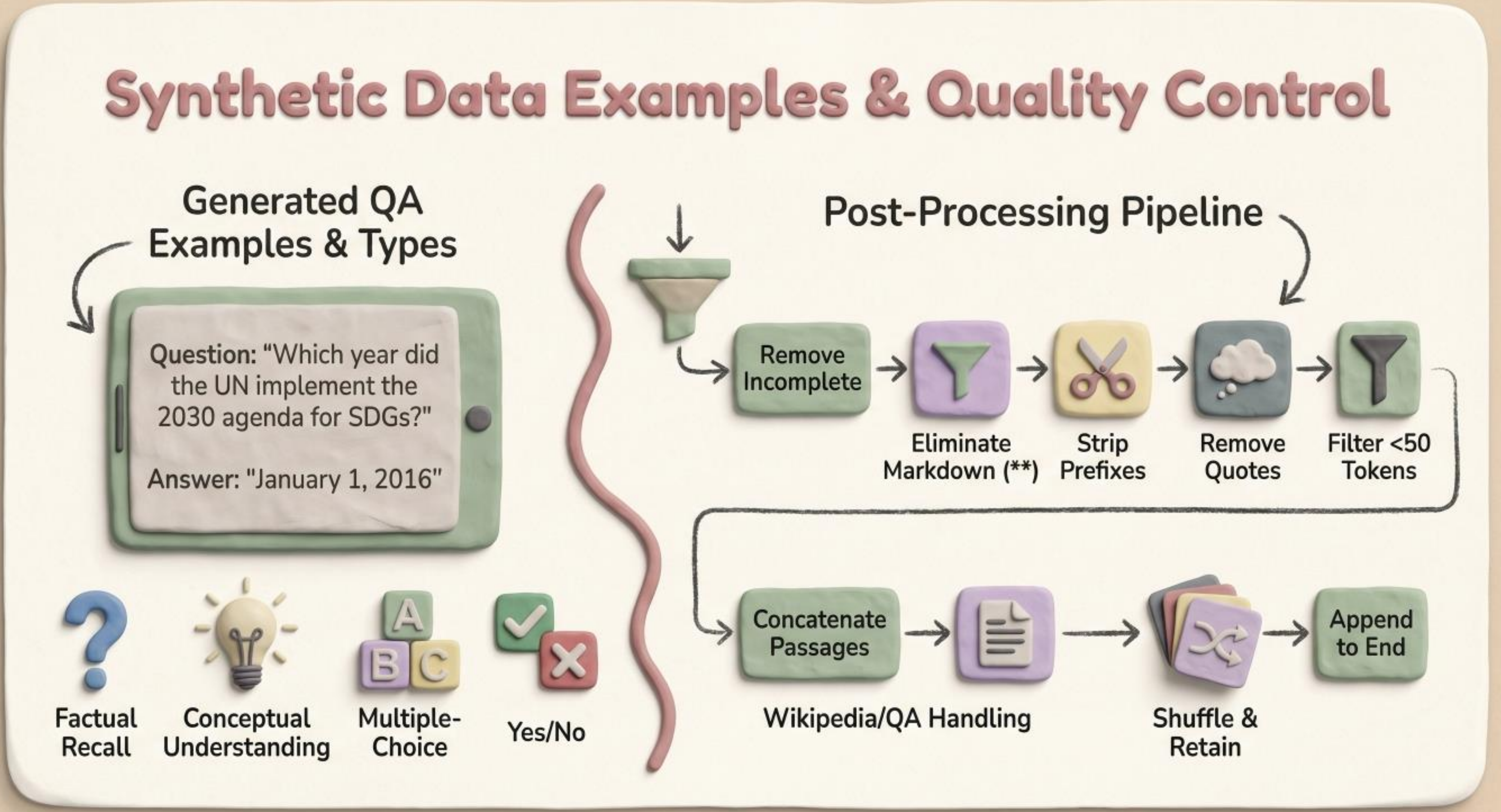
Model-Based Quality Labeling: Classifier Ensemble

Overcoming Single Classifier Bottlenecks with a Three-Pronged Approach

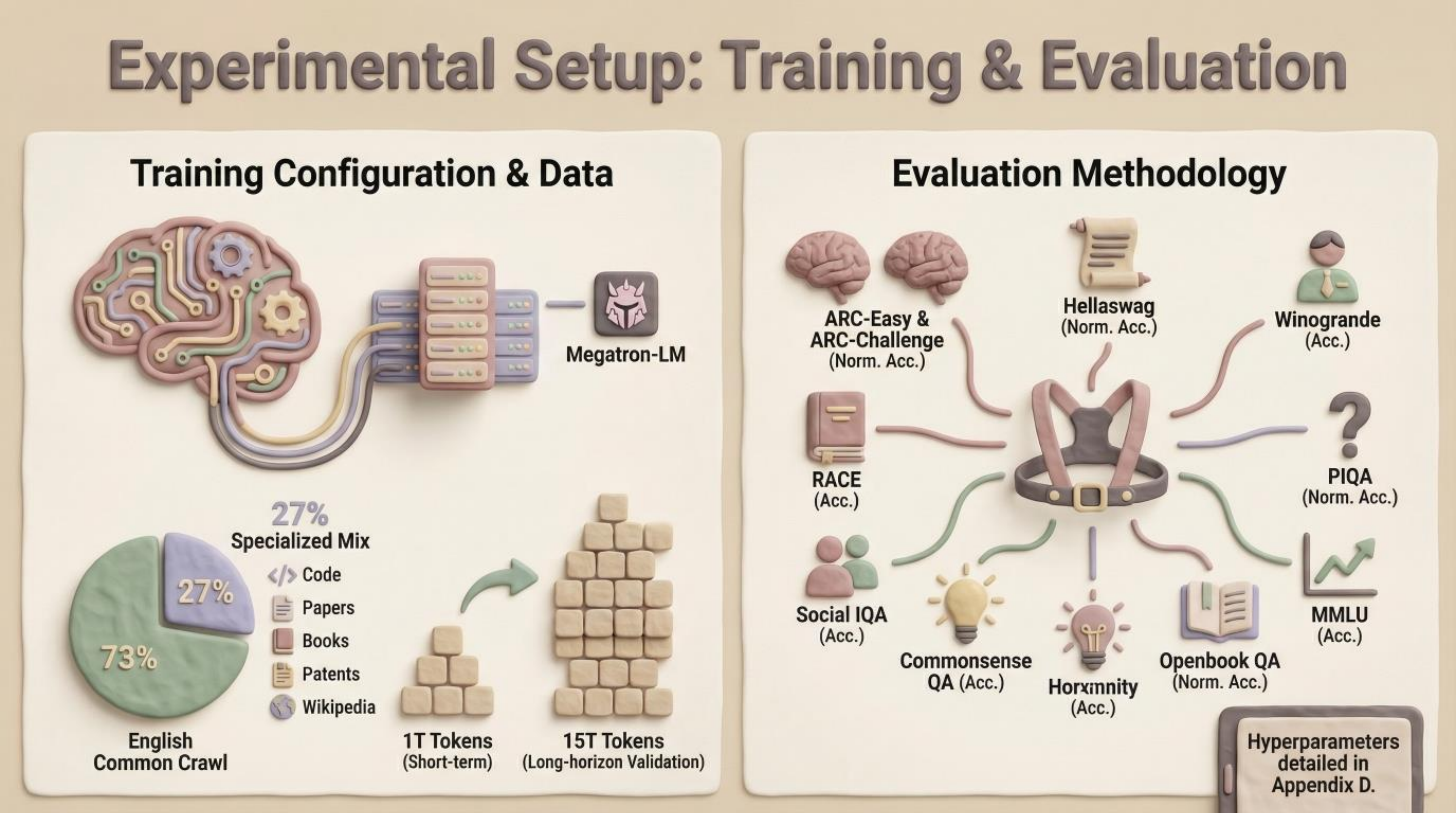



Quality Bucketing: From Scores to Training Labels
20 Fine-Grained Buckets (0-19)
Each ~5%
50B Tokens
8B Model
9 Tasks Avg.
66% Default
34% Evaluated Bucket
High
(Bucket 19, 553B, 12.63%)
Medium-High
(Bucket 18, 504B, 11.52%)
Medium
(Buckets 12-17, 2,023B, 46.24%)
Medium-Low
(Buckets 7-11, 894B, 20.43%)
Low
(Buckets 0-6, 402B, 9.18%)
Aligned with Downstream Performance
Bucketing Methodology
Annealing Evaluation (Regrouping Process)
Final 5-Category Grouping
NVIDIA Research

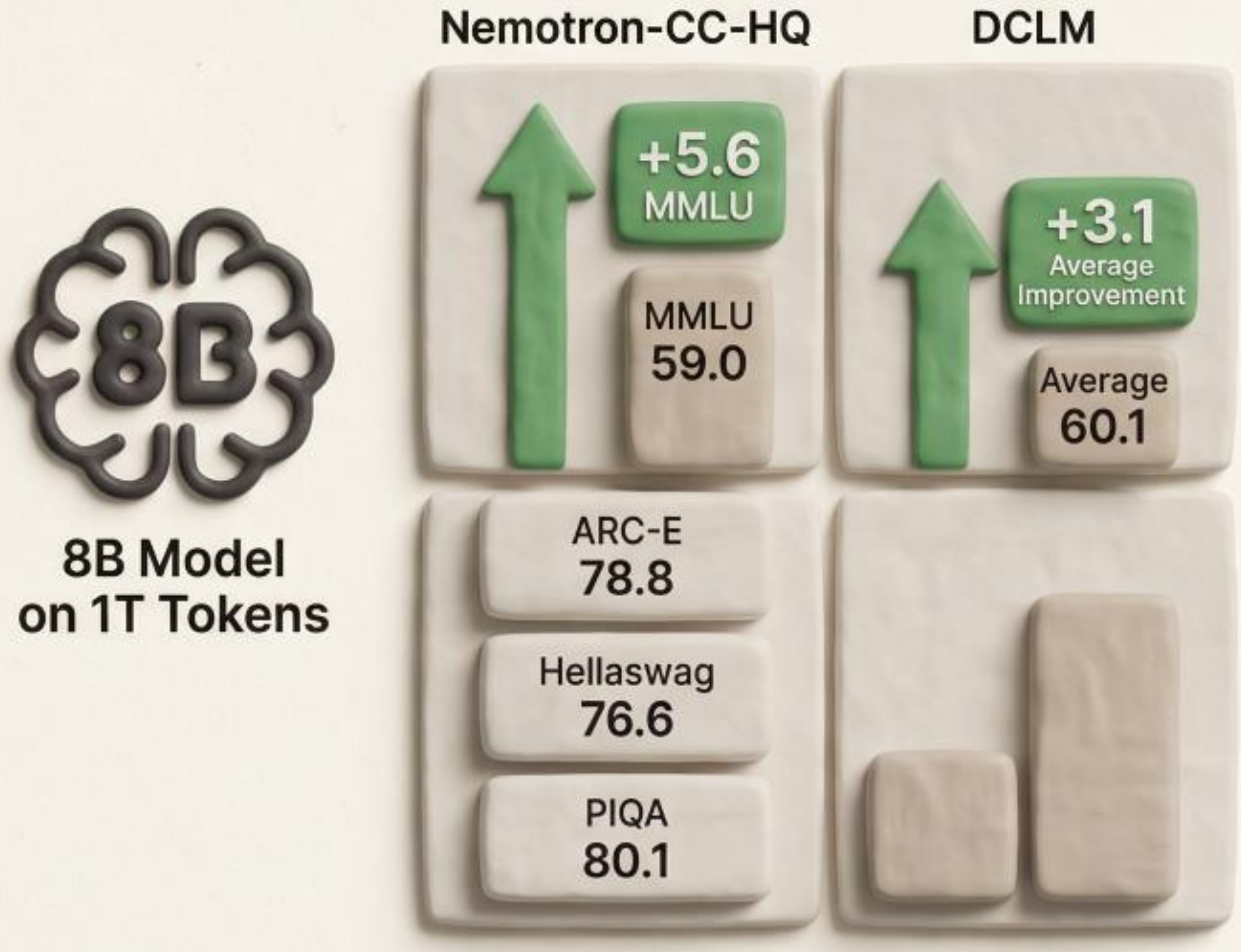

Synthetic Data Generation: Dual Strategies
Raw Data
Low-Quality Data
High-Quality Data
Mistral NeMo 12B
Wikipedia-style Rephrasing
• Goal: Reduce noise/errors, preserve info
• Result: 336.3B Tokens (from 403B raw)
(1) Diverse QA Pairs
?
?
?
A
499.5B Tokens
(2) Distill
157.6B Tokens
(3) Extract Knowledge
303.6B Tokens
(4) Knowledge List
203.2B Tokens
Post-processing & Filtering
• Remove incomplete results
• Strip Markdown
• Filter under-length (<50 tokens)
Refined Synthetic Dataset

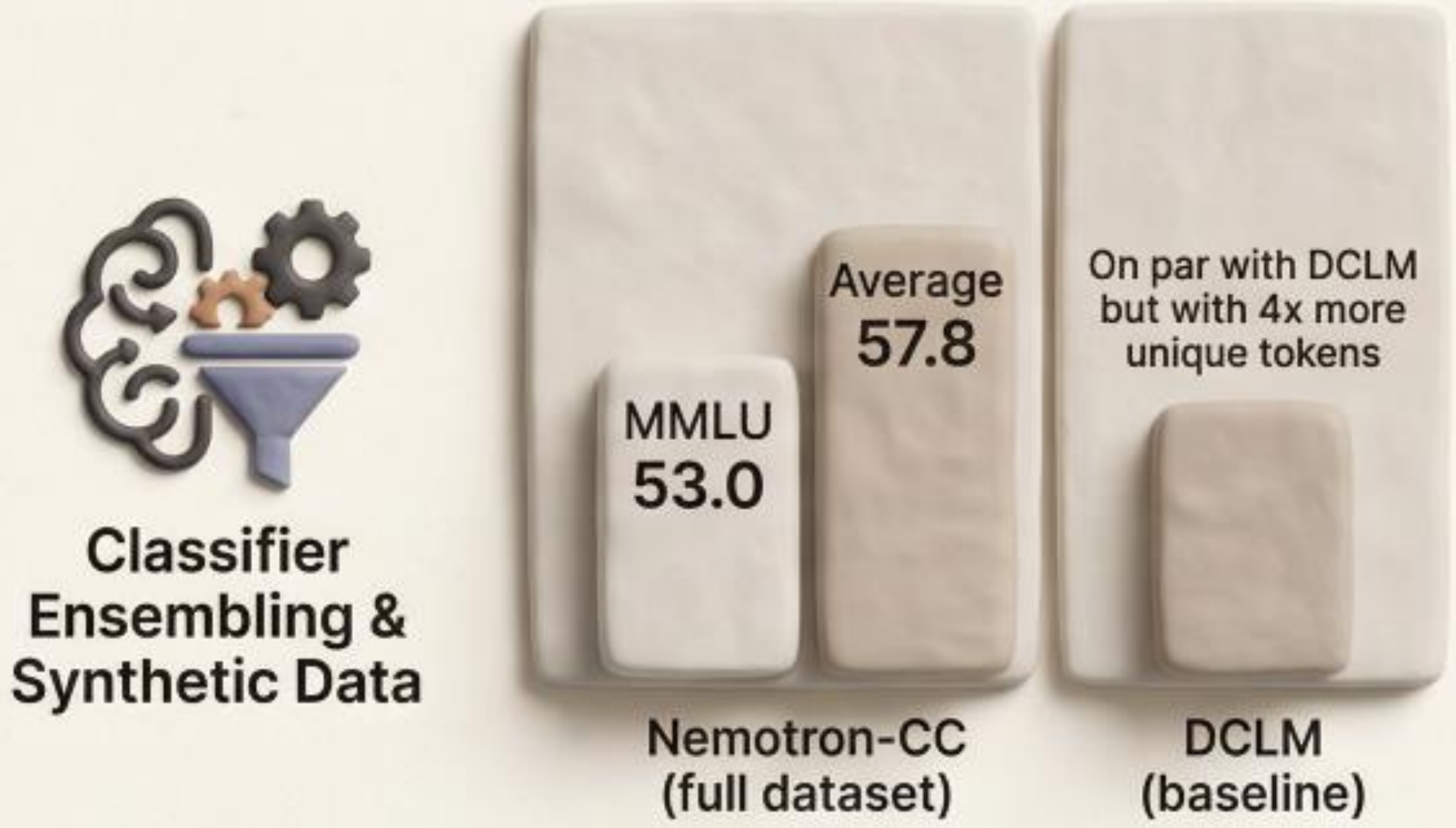


Synthetic Data Examples & Quality Control
Generated QA Examples & Types
Question: "Which year did the UN implement the 2030 agenda for SDGs?"
Answer: "January 1, 2016"
Factual Recall
Conceptual Understanding
A
B C
Multiple-Choice
Yes/No
Post-Processing Pipeline
Remove Incomplete
Eliminate Markdown (**)
Strip Prefixes
Remove Quotes
Filter <50 Tokens
Concatenate Passages
Wikipedia/QA Handling
Shuffle & Retain
Append to End


Experimental Setup: Training & Evaluation
Training Configuration & Data
Megatron-LM
27%
Specialized Mix
</> Code
Papers
Books
Patents
Wikipedia
27%
73%
English
Common Crawl
1T Tokens
(Short-term)
15T Tokens
(Long-horizon Validation)
Evaluation Methodology
ARC-Easy &
ARC-Challenge
(Norm. Acc.)
Hellaswag
(Norm. Acc.)
Winogrande
(Acc.)
RACE
(Acc.)
PIQA
(Norm. Acc.)
Social IQA
(Acc.)
Commonsense
QA (Acc.)
Horxmnity
(Acc.)
Openbook QA
(Norm. Acc.)
MMLU
(Acc.)
Hyperparameters
detailed in
Appendix D.

Short Token Horizon Results (1T Tokens)





Key Insight: Classifier ensembling and synthetic data effective even in non-data-constrained settings. Superiority demonstrated across most tasks.

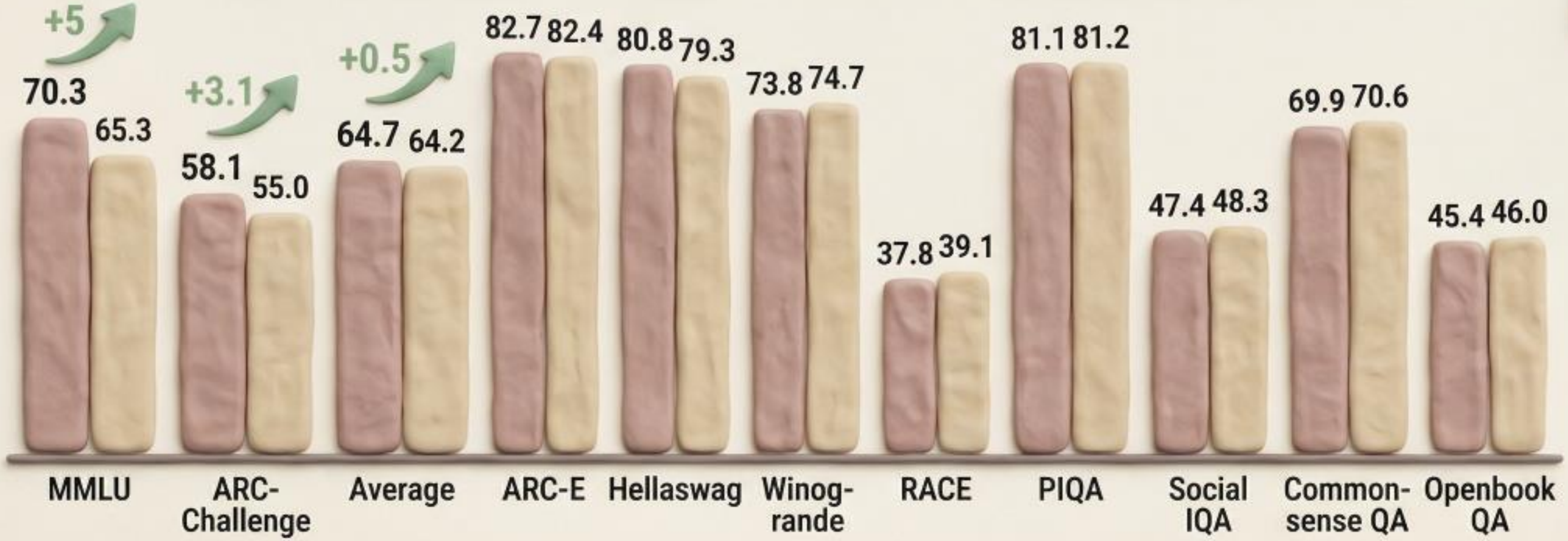
Long Token Horizon Results (15T Tokens)

Key Achievements

+5 MMLU

+3.1 ARC-Challenge

+0.5 Average



Validates dataset suitability for state-of-the-art long-horizon training.

Note: Evaluation uses lm-evaluation-harness; Meta's numbers may differ due to customizations.

NVIDIA Research

Ablation Study: Extractor & Filter Impact

Configurations & Performance

Configuration	MMLU	Average
Trafilatura filtered	55.4	60.6
Justext filtered	54.1	60.9
Justext unfiltered	55.5	60.3
Justext HQ-unfiltered	57.5	60.6

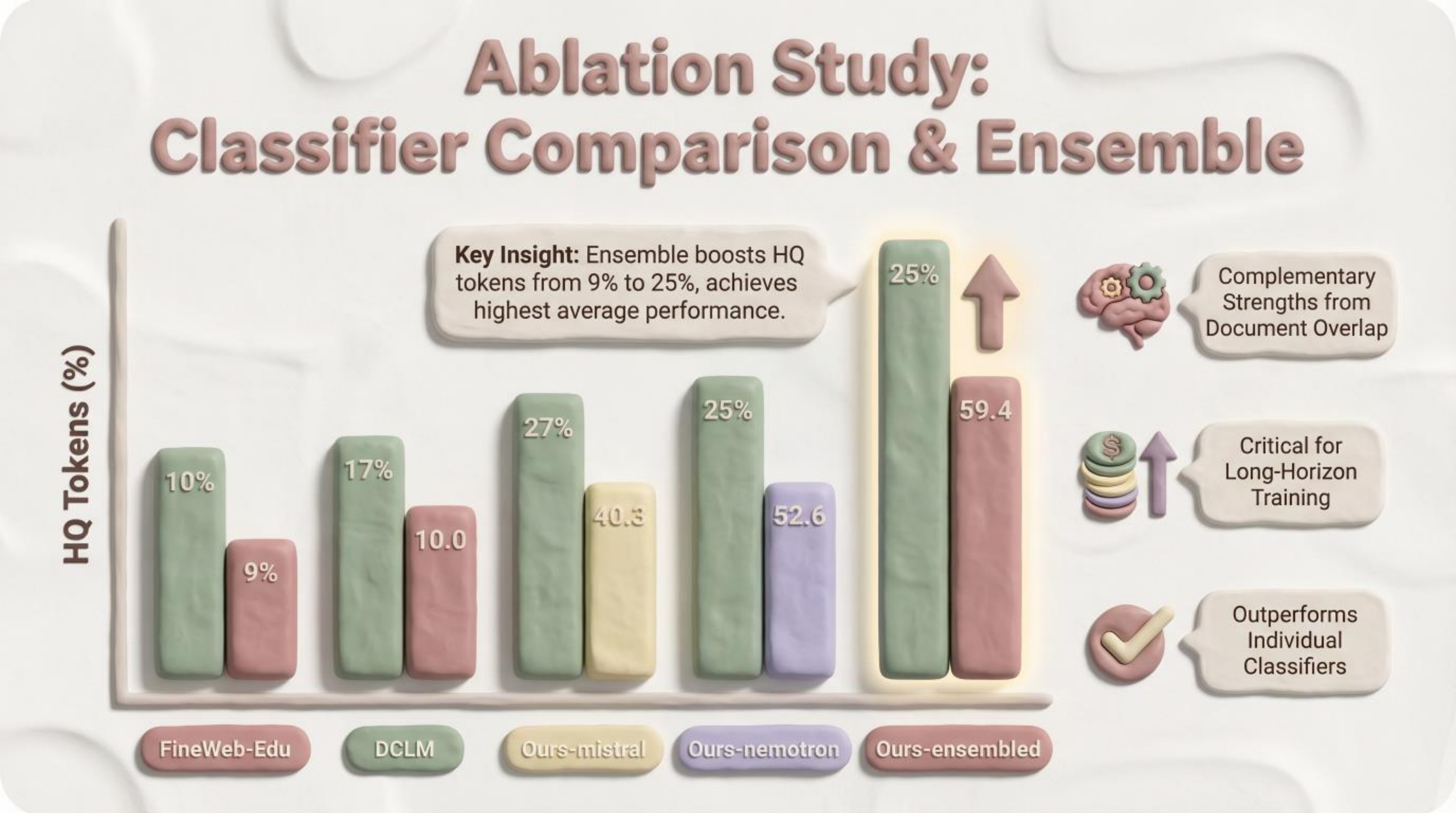
Finn · Com · Best

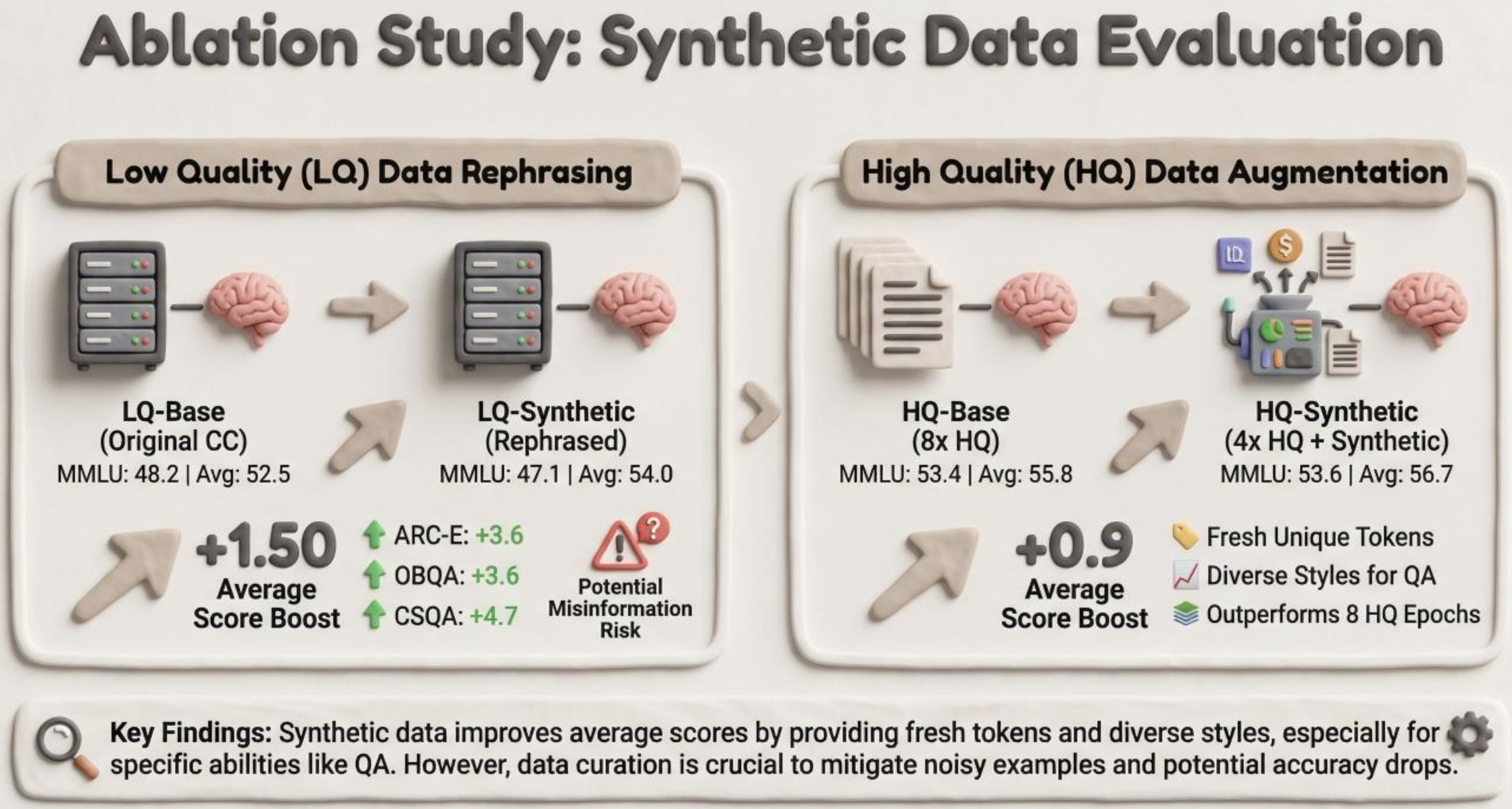
Key Findings & Strategic Insight

- Justext yields 57.4% more HQ tokens than Trafilatura without accuracy impact
- Removing filters from HQ tokens improves MMLU by +2%
- Combining Justext + no HQ filtering maximizes both token yield and accuracy

+2%

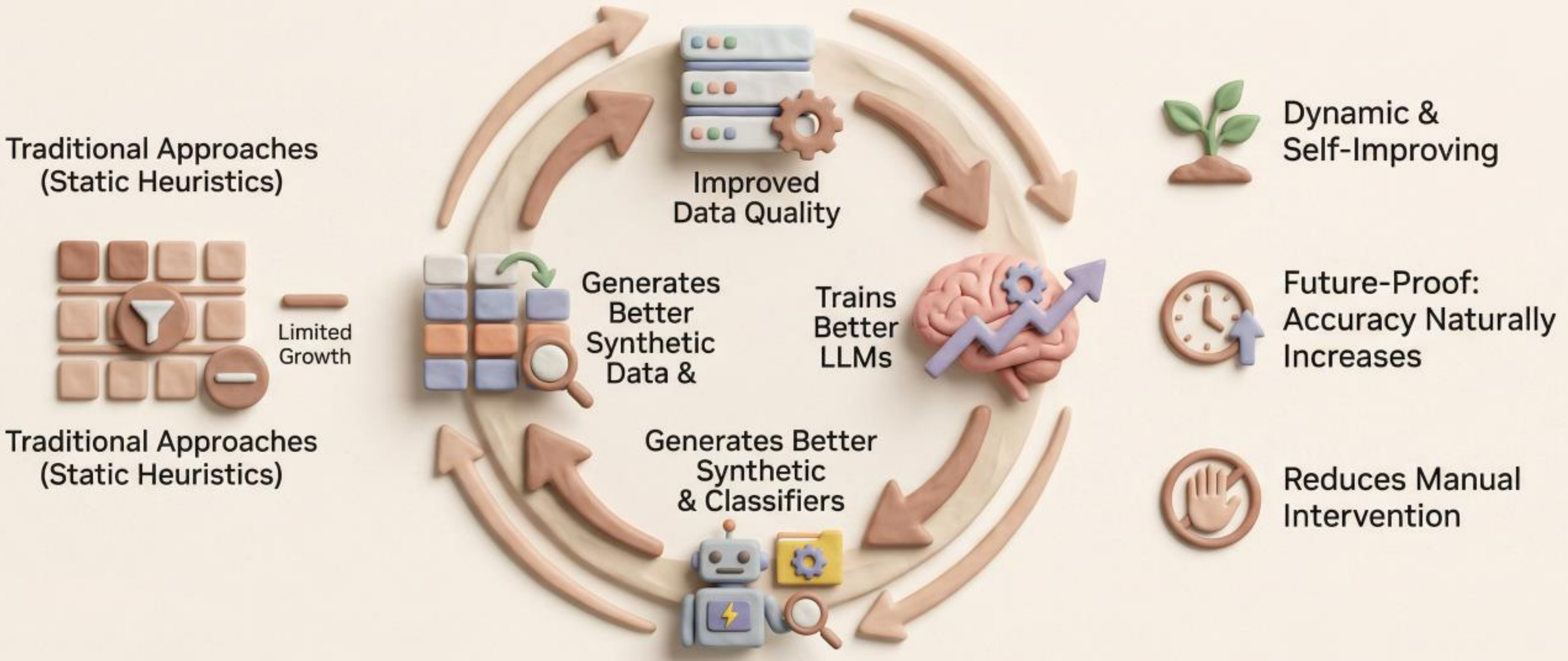
Strategic Insight: Filters remove 18.1% of HQ tokens unnecessarily.
Conclusion: Apply heuristics only to low-quality tokens.


Ablation Study:
Classifier Comparison & Ensemble
HQ Tokens (%)
Key Insight: Ensemble boosts HQ tokens from 9% to 25%, achieves highest average performance.
10%
9%
17%
10.0
27%
40.3
25%
52.6
25%
59.4
FineWeb-Edu
DCLM
Ours-mistral
Ours-nemotron
Ours-ensembled
Complementary Strengths from Document Overlap
Critical for Long-Horizon Training
Outperforms Individual Classifiers

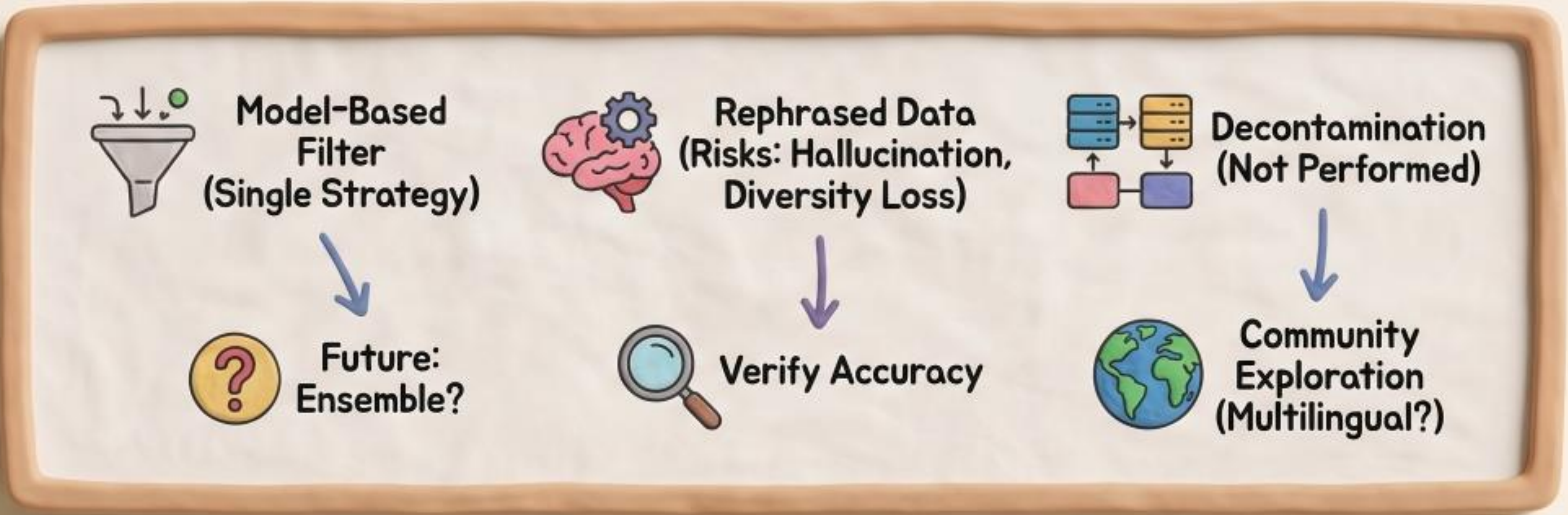


Ablation Study: Synthetic Data Evaluation
Low Quality (LQ) Data Rephrasing
LQ-Base
(Original CC)
MMLU: 48.2 | Avg: 52.5
LQ-Synthetic
(Rephrased)
MMLU: 47.1 | Avg: 54.0
+1.50
Average
Score Boost
ARC-E: +3.6
OBQA: +3.6
CSQA: +4.7
Potential
Misinformation
Risk
High Quality (HQ) Data Augmentation
HQ-Base
(8x HQ)
MMLU: 53.4 | Avg: 55.8
HQ-Synthetic
(4x HQ + Synthetic)
MMLU: 53.6 | Avg: 56.7
+0.9
Average
Score Boost
Fresh Unique Tokens
Diverse Styles for QA
Outperforms 8 HQ Epochs
Key Findings: Synthetic data improves average scores by providing fresh tokens and diverse styles, especially for specific abilities like QA. However, data curation is crucial to mitigate noisy examples and potential accuracy drops.

Guiding Principle: The Learned Flywheel



Limitations & Future Directions



Filtering & Pipeline

- Single filter strategy tested; needs ensemble for higher quality.
- Pipeline components (e.g., lang ID) not fully ablated.

Rephrasing & Scope

- Factual accuracy & fidelity of rephrased data unverified.
- Tested on English only; multilingual extension needed.

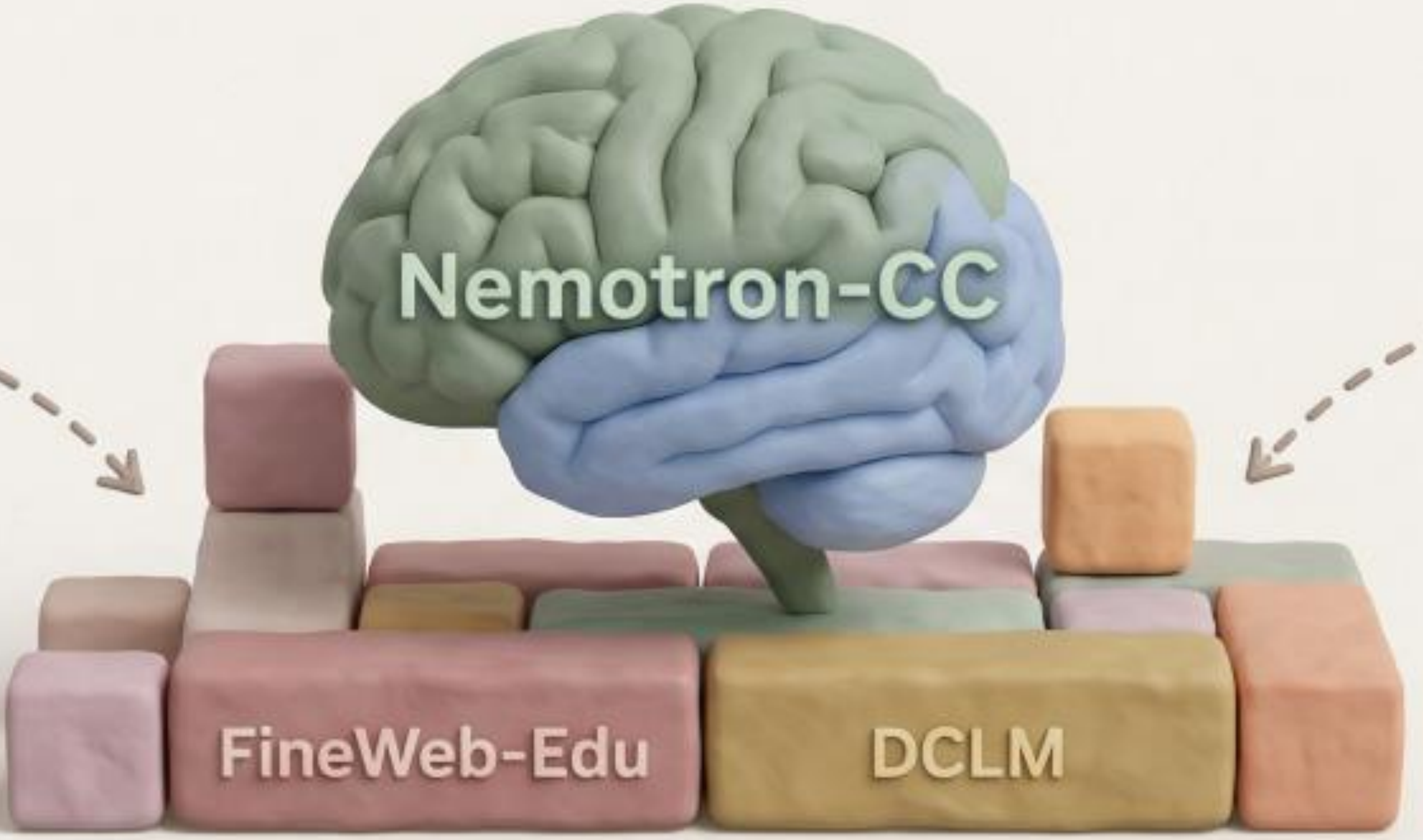
Decontamination

- Dataset not decontaminated due to lack of consensus.
- Requires community exploration on best practices.

Related Work: Building on Previous Research

Phi Series

Pioneered high-quality small data; limited details/horizon



Other CC Datasets (C4, DOLMA, etc.)

Focus on extraction & heuristics; Nemotron-CC goes beyond

Builds on model-based filtering, improves quantity (ensemble, rephrasing, reduced heuristics)

Multilingual Datasets

English focus enables depth; multilingual extension future work

Concurrent work combining existing datasets; Nemotron-CC creates NEW from raw CC

Synthetic Data Literature & Inspirations

Instruction Pre-training
(Cheng et al. 2024)

Synthesized instruction-response pairs for pre-training.

TinyStories
(Eldan and Li 2023)

Small models trained on **synthetic short stories** generate fluent narratives.

Textbook Models
(Gunasekar et al. 2023)

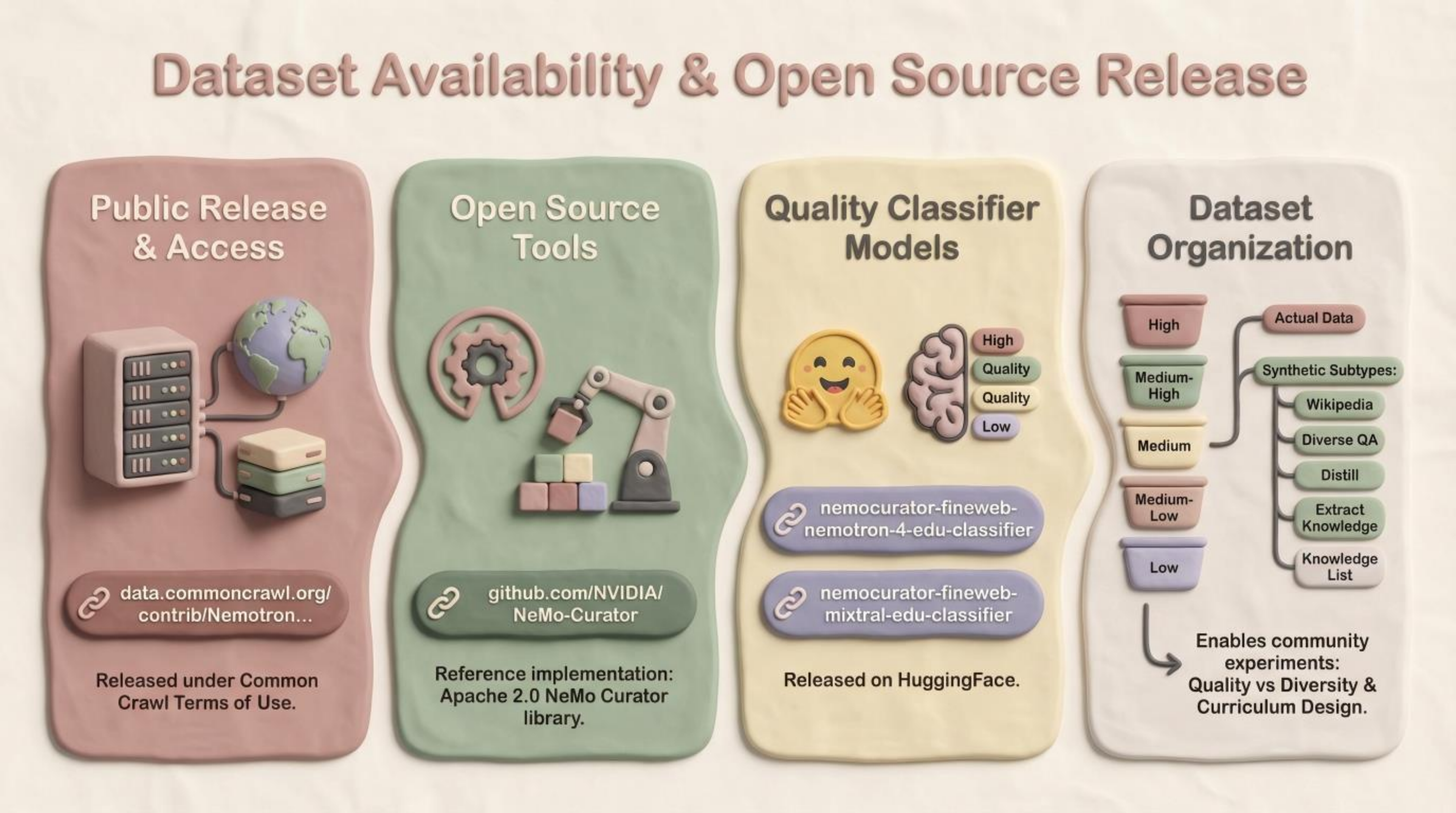
Synthetic textbooks and exercises achieve impressive coding benchmarks.

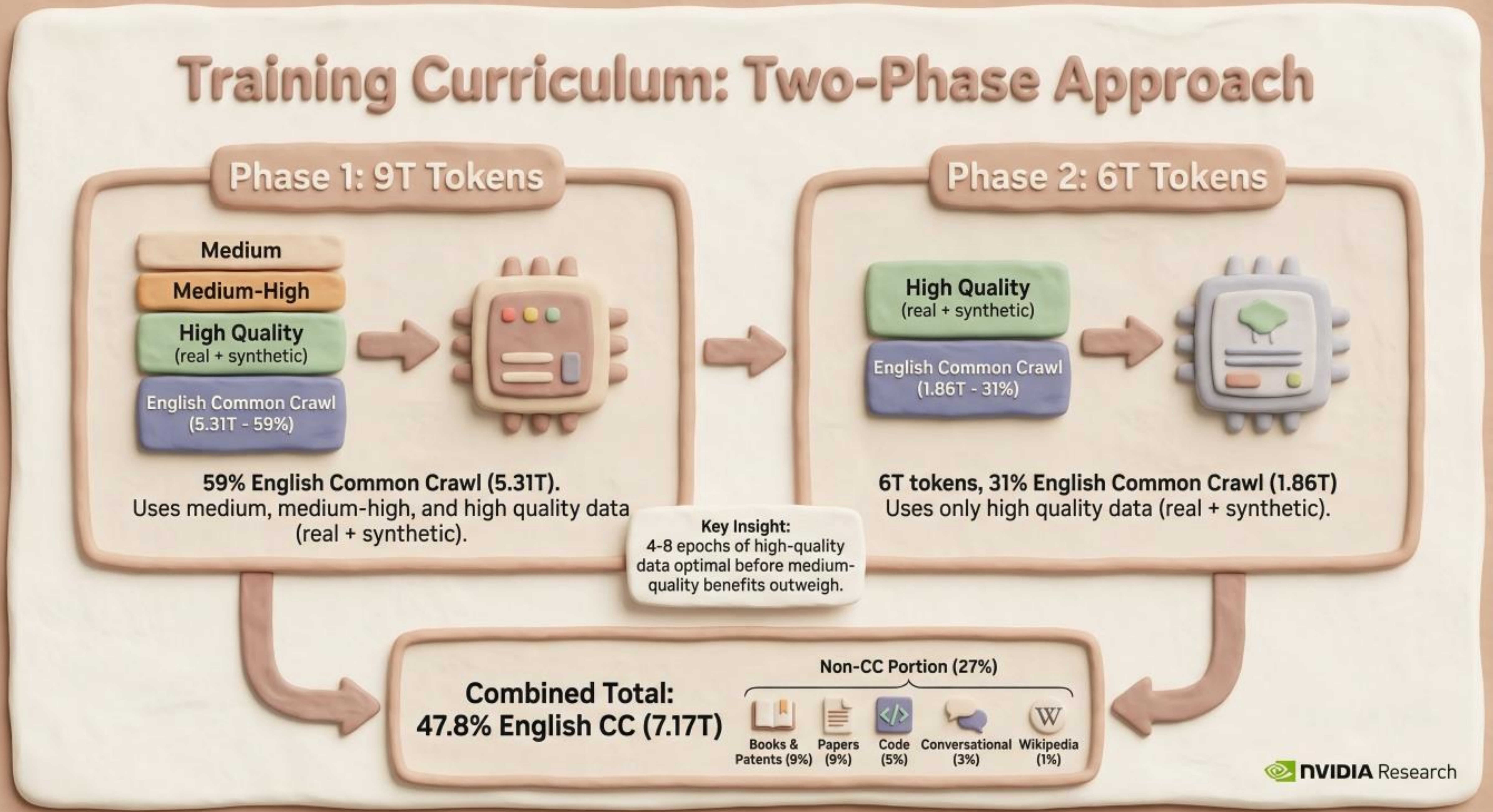
Rephrasing Approach (Maini et al. 2024)

Smaller models (Qwen-1.8B, Mistral-7B) adequate for web data rephrasing.

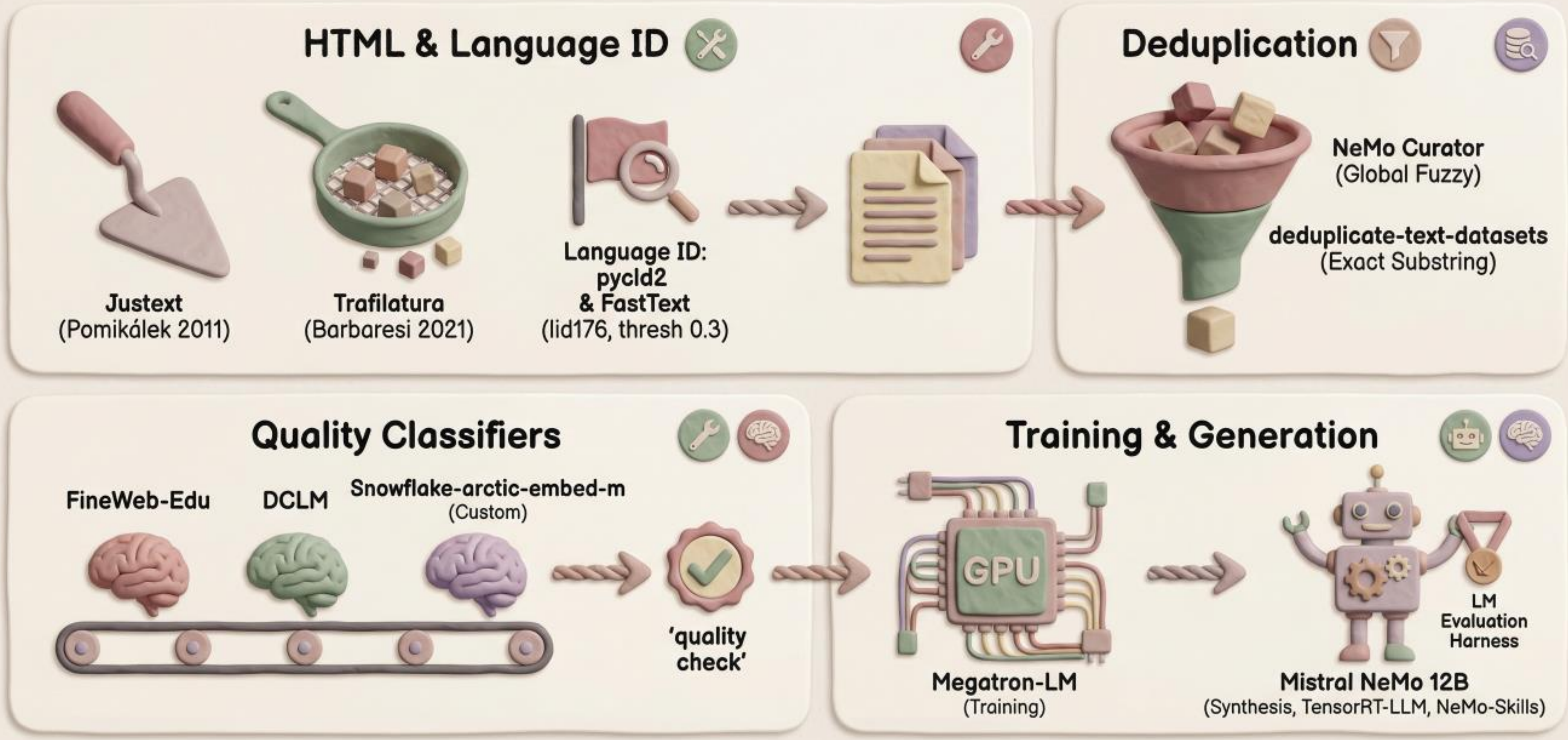
Nemotron-CC Extension

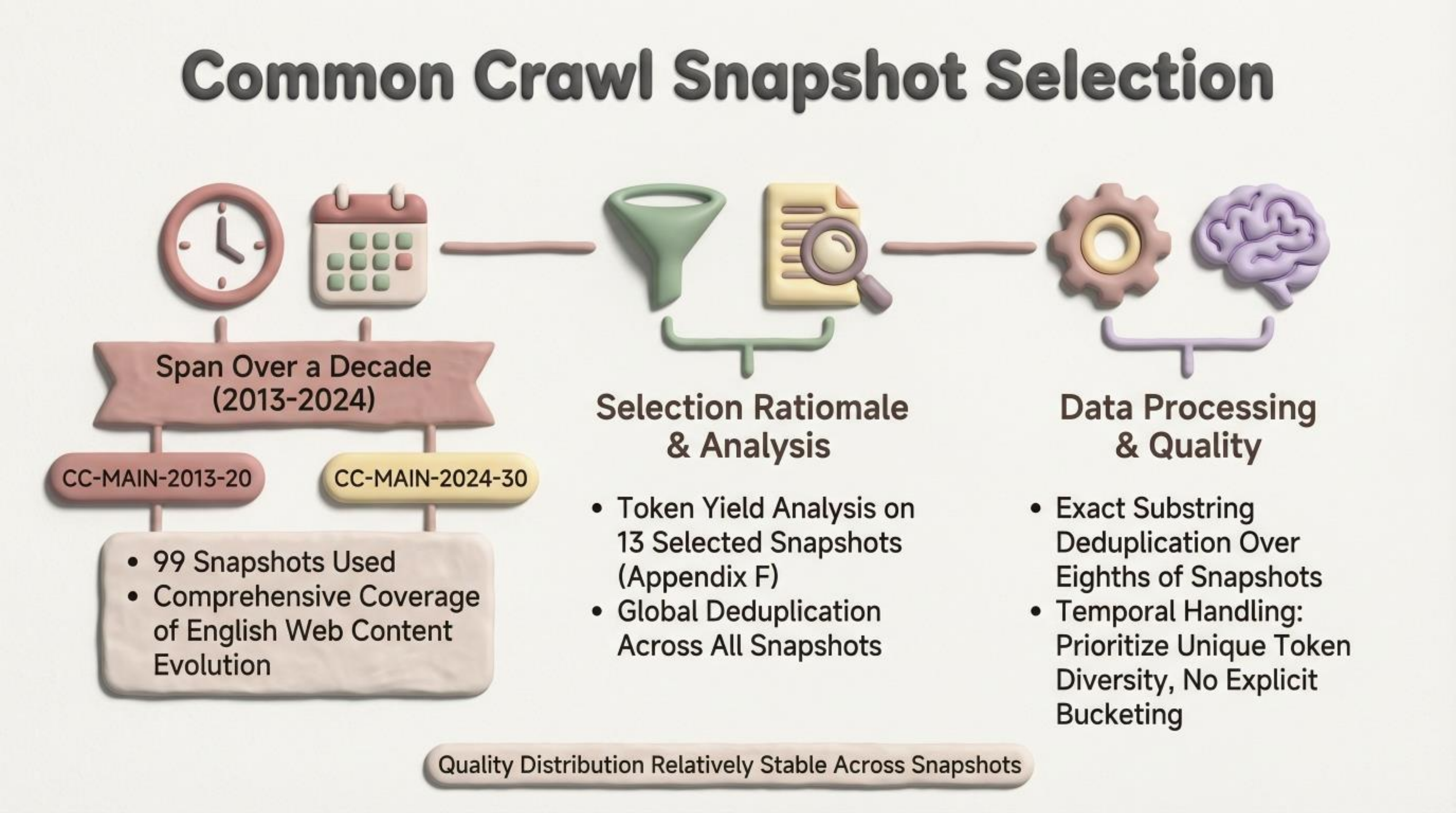
Adopts and extends: more prompts, specialized for quality tiers, large-scale demonstration


Dataset Availability & Open Source Release
Public Release & Access
data.commoncrawl.org/contrib/Nemotron...
Released under Common Crawl Terms of Use.
Open Source Tools
github.com/NVIDIA/NeMo-Curator
Reference implementation: Apache 2.0 NeMo Curator library.
Quality Classifier Models
High
Quality
Quality
Low
nemocurator-fineweb-nemotron-4-edu-classifier
nemocurator-fineweb-mixtral-edu-classifier
Released on HuggingFace.
Dataset Organization
High
Medium-High
Medium
Medium-Low
Low
Actual Data
Synthetic Subtypes:
Wikipedia
Diverse QA
Distill
Extract Knowledge
Knowledge List
Enables community experiments: Quality vs Diversity & Curriculum Design.


Training Curriculum: Two-Phase Approach
Phase 1: 9T Tokens
Medium
Medium-High
High Quality
(real + synthetic)
English Common Crawl
(5.31T - 59%)
59% English Common Crawl (5.31T).
Uses medium, medium-high, and high quality data (real + synthetic).
Phase 2: 6T Tokens
High Quality
(real + synthetic)
English Common Crawl
(1.86T - 31%)
6T tokens, 31% English Common Crawl (1.86T)
Uses only high quality data (real + synthetic).
Key Insight:
4-8 epochs of high-quality data optimal before medium-quality benefits outweigh.
Combined Total:
47.8% English CC (7.17T)
Non-CC Portion (27%)
Books & Patents (9%)
Papers (9%)
Code (5%)
Conversational (3%)
Wikipedia (1%)
NVIDIA Research

Technical Implementation: Tools & Infrastructure




Common Crawl Snapshot Selection
Span Over a Decade (2013-2024)
CC-MAIN-2013-20
CC-MAIN-2024-30
• 99 Snapshots Used
• Comprehensive Coverage of English Web Content Evolution
Selection Ratiomale & Analysis
• Token Yield Analysis on 13 Selected Snapshots (Appendix F)
• Global Deduplication Across All Snapshots
Data Processing & Quality
• Exact Substring Deduplication Over Eighths of Snapshots
• Temporal Handling: Prioritize Unique Token Diversity, No Explicit Bucketing
Quality Distribution Relatively Stable Across Snapshots

Quality Control: Deduplication Strategy

Comprehensive Approach

1. Global Exact
   - Remove identical docs across 99 snapshots
2. Global Fuzzy
   - Near-duplicate detection (similarity hashing). Tool: NeMo Curator
3. Exact Substring
   - Split chunks into 8 parts, deduplicate within. Tool: deduplicate-text-datasets

Rationale & Impact

- Diminishing returns after 4 epochs (Muennighoff et al., 2024)

- Enables longer effective training without repetition penalty
- Trade-off: Some near-duplicates may preserve style

Evaluation Metrics: Comprehensive Benchmark Suite

Reasoning & Commonsense

- ARC-Easy & ARC-Challenge
- Hellaswag
- Winogrande
- PIQA
- Social IQA
- Commonsense QA
- Openbook QA

Reading Comprehension

- RACE (middle/high school exams)

Knowledge & Reasoning

- MMLU (57 subjects across STEM, humanities, social sciences)

Metrics & Coverage

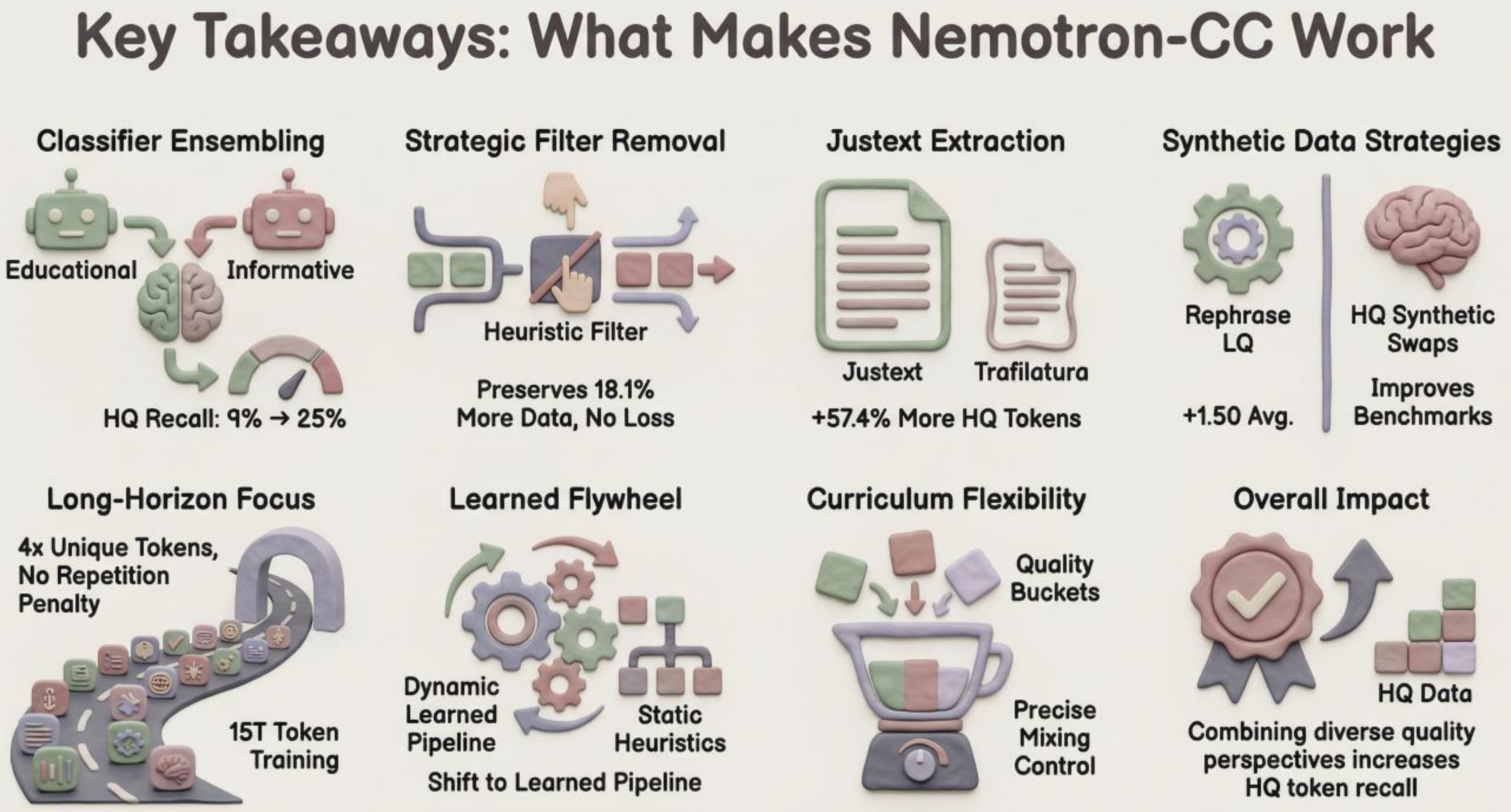
Normalized & Raw Accuracy

Science, Physical, Social, Reading, Broad Knowledge

Standardized Evaluation

lm-evaluation-harness for fair comparison

Reported numbers may differ due to implementation.


Key Takeaways: What Makes Nemotron-CC Work
Classifier Ensembling
Educational
Informative
HQ Recall: 9% → 25%
Strategic Filter Removal
Heuristic Filter
Preserves 18.1%
More Data, No Loss
Justext Extraction
Justext
Trafilatura
+57.4% More HQ Tokens
Synthetic Data Strategies
Rephrase LQ
+1.50 Avg.
HQ Synthetic Swaps
Improves Benchmarks
Long-Horizon Focus
4x Unique Tokens, No Repetition Penalty
15T Token Training
Learned Flywheel
Dynamic Learned Pipeline
Static Heuristics
Shift to Learned Pipeline
Curriculum Flexibility
Quality Buckets
Precise Mixing Control
Overall Impact
HQ Data
Combining diverse quality perspectives increases HQ token recall

Impact on the LLM Community

RESEARCH & INDUSTRY IMPLICATIONS

METHODOLOGICAL CONTRIBUTIONS & OPEN SCIENCE

- Classifier Ensembling, Strategic Filtering
- Full Pipeline, Classifiers, Ablations Released for Reproducibility
- Foundation for Future Research: Quality-Bucketed Structure

DEMOCRATIZING PRETRAINING

- Public 6.3T Token Dataset (Proven Quality)
- Enabling Smaller Labs: Competitive Models without Massive Proprietary Data
  - 15T+ Token Training Feasible for 8B+ Models

NEW QUALITY-QUANTITY BALANCE

- Challenging Aggressive Filtering Assumption
- Quality-Aware Synthesis as Reusable Technique

INDUSTRY IMPACT & INNOVATION

- Better Open Models Drive Competition
- Long-Horizon Training Unlocked

Comprehensive Overview for University Lecture

Conclusion: Advancing Long-Horizon Pretraining