

NVIDIA Nemotron Nano 2: Technical Deep Dive



An Accurate and Efficient Hybrid Mamba-Transformer Reasoning Model

Model Overview & Key Achievements



Core Identity

- Hybrid Mamba-Transformer Model
- Achieving 3-6x Higher Throughput vs. Qwen3-8B



Key Metrics

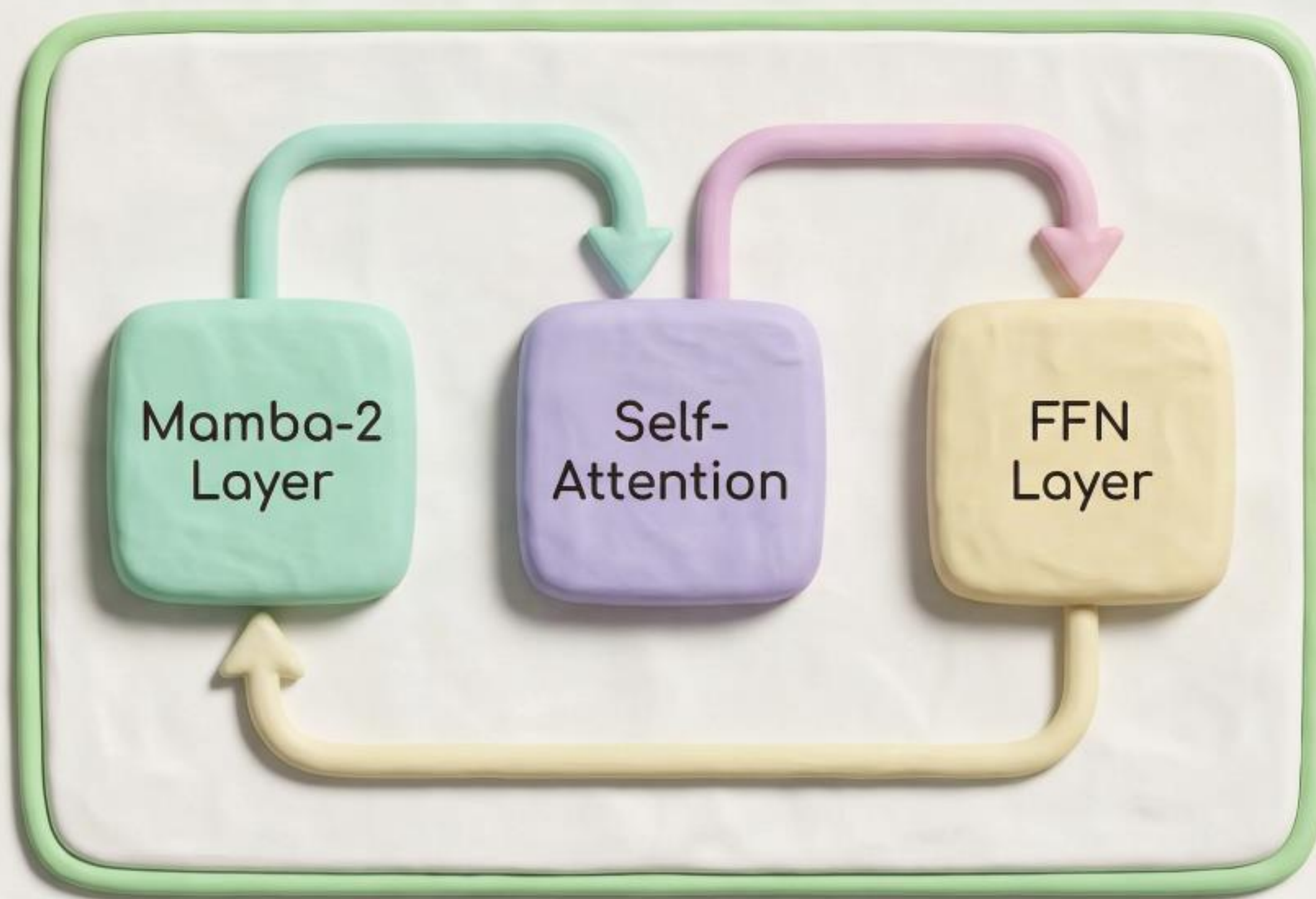
- 9B Parameters
- 128k Context Length
- 20T Token Training



Performance

- Competitive Accuracy
- Dramatically Higher Inference Speed

Model Architecture: Hybrid Design

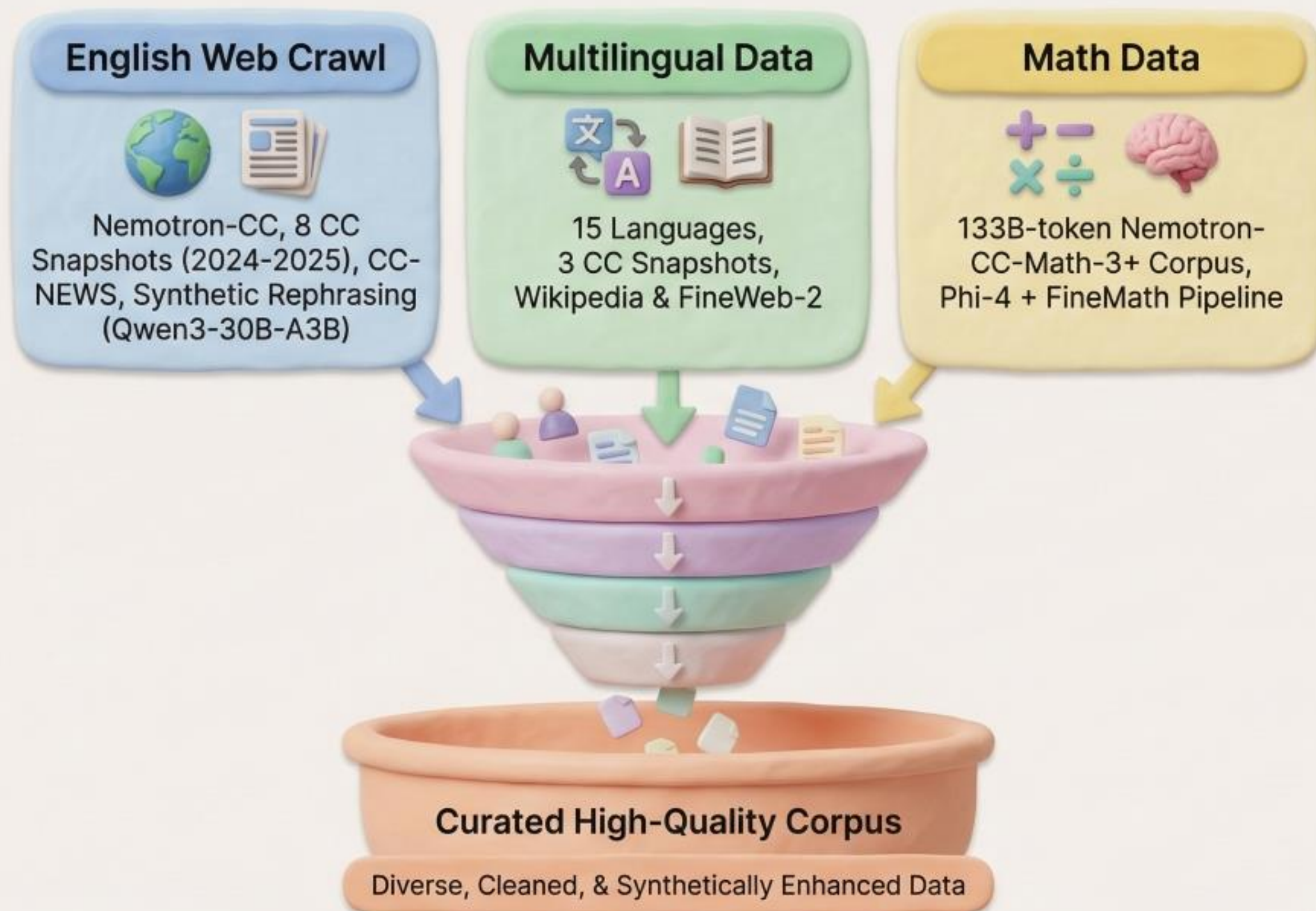


- 62 Layers (6 Attention, 28 FFN, 28 Mamba-2)
- 5120 Hidden Dim
- 20480 FFN Dim
- 40 Query Heads
- 8 KV Heads



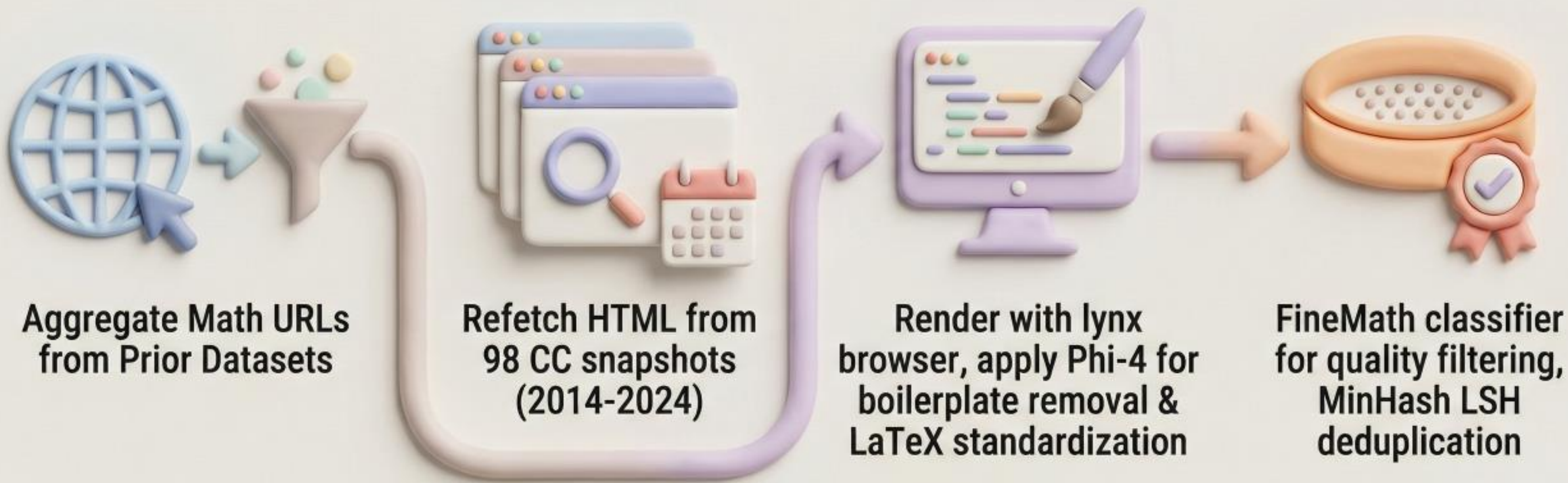
**8% Attention
Layer Ratio**
For speed optimization

Pre-Training Data: Curated Sources



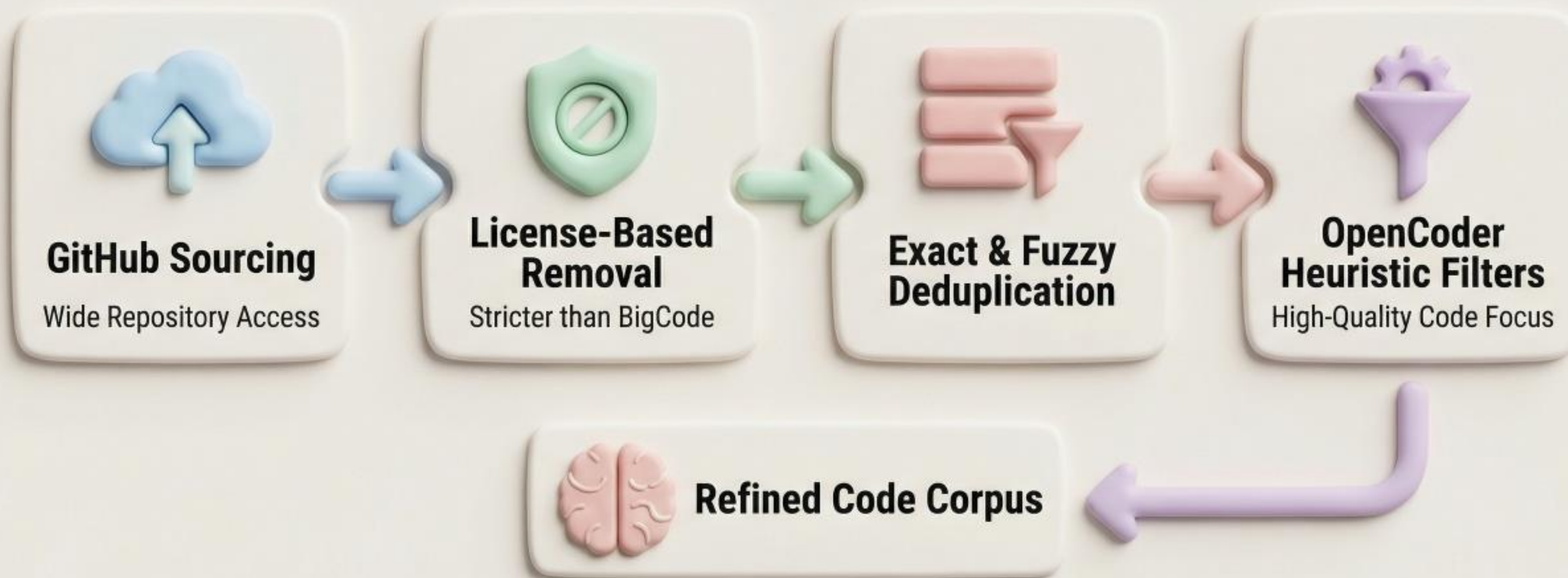
Pre-Training Data: Math Extraction Pipeline

Overview: Specialized Math Extraction Process



Pre-Training Data: Code & Synthetics

Code Data Processing Pipeline



Data Mixture & Training Strategy

Phase 1: Diversity-Focused (0-60%)

Crawl-Medium	18.3%
Crawl-Medium-High	14.8%
Crawl-High	11.1%
Syn-Crawl-High	16.2%

Broad data collection
for diversity

Phase 2: Quality-Focused (60-90%)

Wikipedia	0.9%
Crawl-High	16%
Syn-Crawl-High	21%

Refining and prioritizing
high-quality sources

Phase 3: Highest Quality (90-100%)

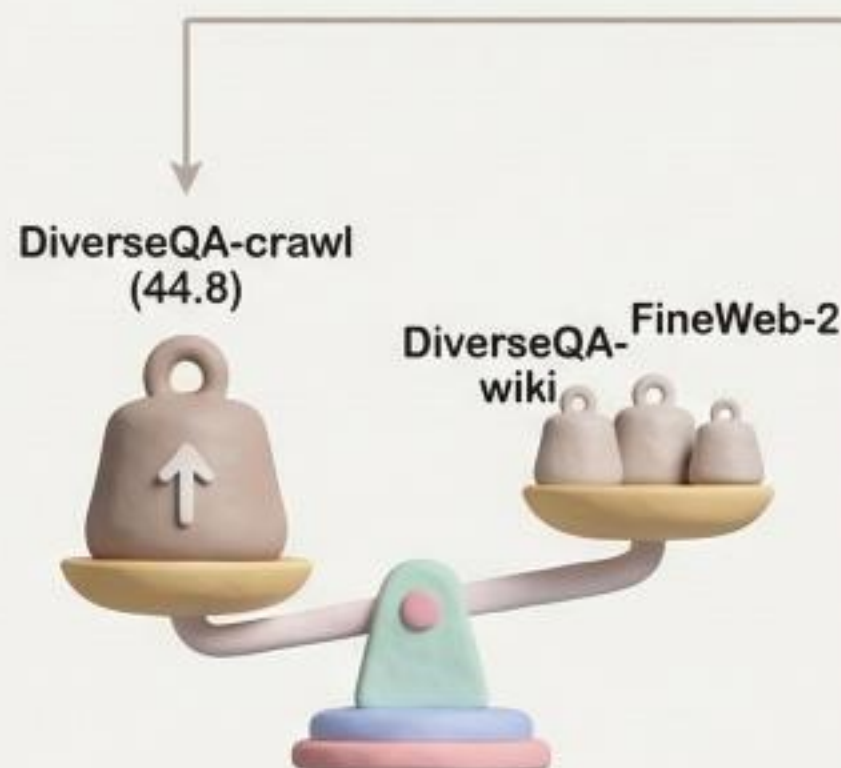
STEM-SFT	32%
Crawl-High	10%
Math	11%

Focus on specialized,
high-value domains

Multilingual Data Ablation Study

Data Source Comparison (Global-MMLU Avg)

- DiverseQA-crawl (translated English CC): 44.8
- DiverseQA-wiki (synthetic from Wikipedia): 42.1
- Common Crawl curated: 37.0
- FineWeb-2: 35.1

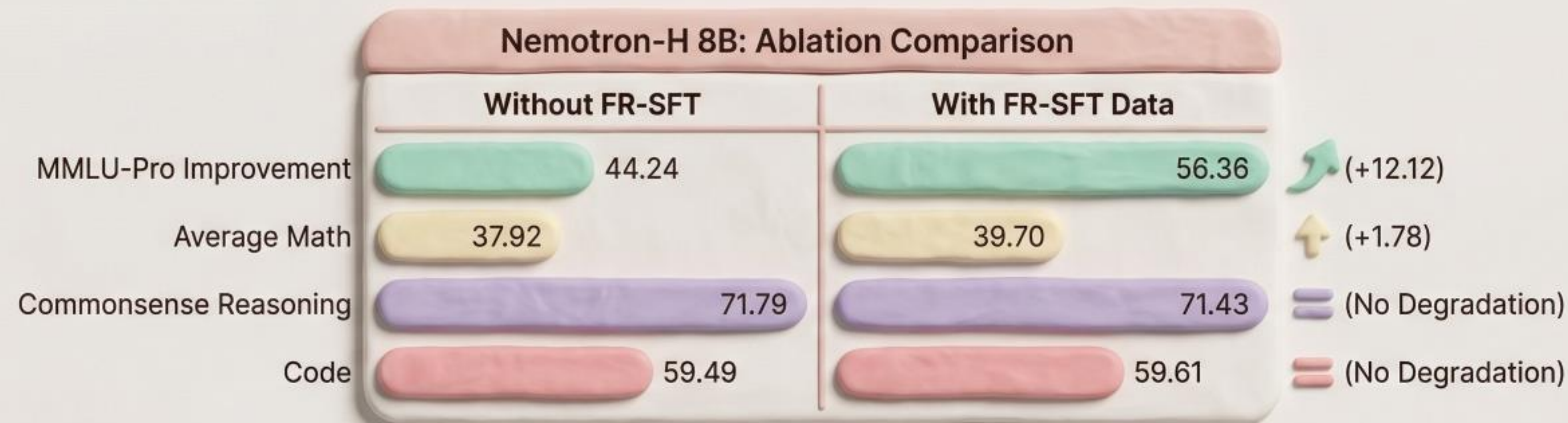


Decision & Rationale

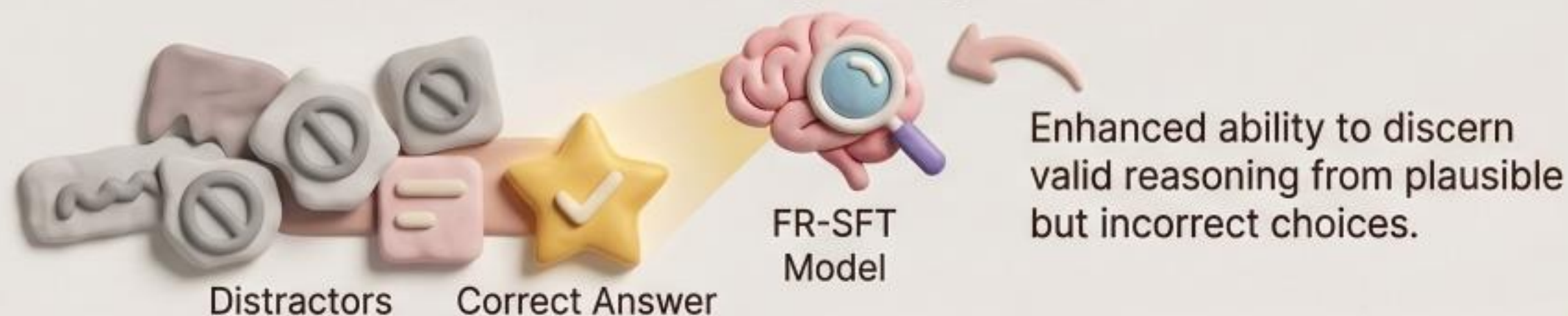
- Prioritize DiverseQA-crawl for training weight.
- Decision based on superior average average performance across 8 languages.
- Ensures highest quality multilingual capabilities.

Fundamental Reasoning SFT Ablation

Table 3 Ablation Results & Mechanism



FR-SFT Mechanism: Distinguishing Answers



Training Configuration & FP8 Recipe

Training Hyperparameters



WSD Learning
Rate Schedule

Stable Phase
4.5e-4



Decay Phase
3.6T Tokens, 4.5e-6 Min Rate

WSD Learning Rate Schedule

8192
Sequence Length

768
Global Batch Size
(6.03M Tokens/Batch)

Adam
 $\beta_1=0.9$, $\beta_2=0.95$, 0.1
Weight Decay

FP8 Recipe



E4M3
Tensors



128x128
Weight Blocks



1x128
Activation
Tiles



FP8 Parameter
All-Gather



FP32
Master
Weights

Long-Context Extension Strategy

Phase LC Continuous Pretraining



 **18.9B**
tokens

 **Constant**
4.5e-6 LR

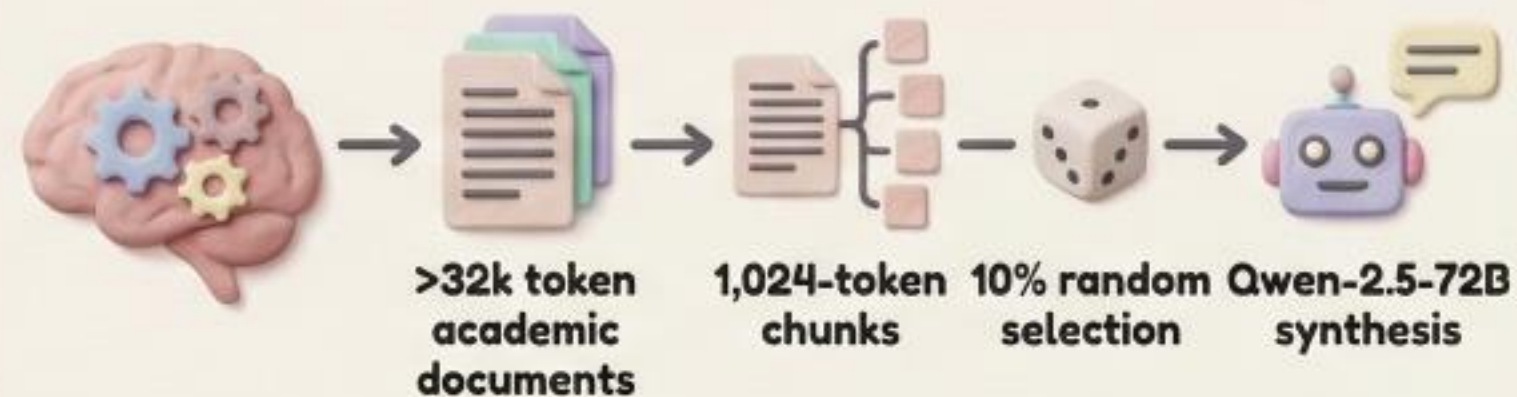


**8-way Tensor +
16-way Context Parallelism**

12_B

**Maintaining 6M
tokens/batch**

Synthetic Long-Context QA Generation



● **Synthetic Long-Context precltation**

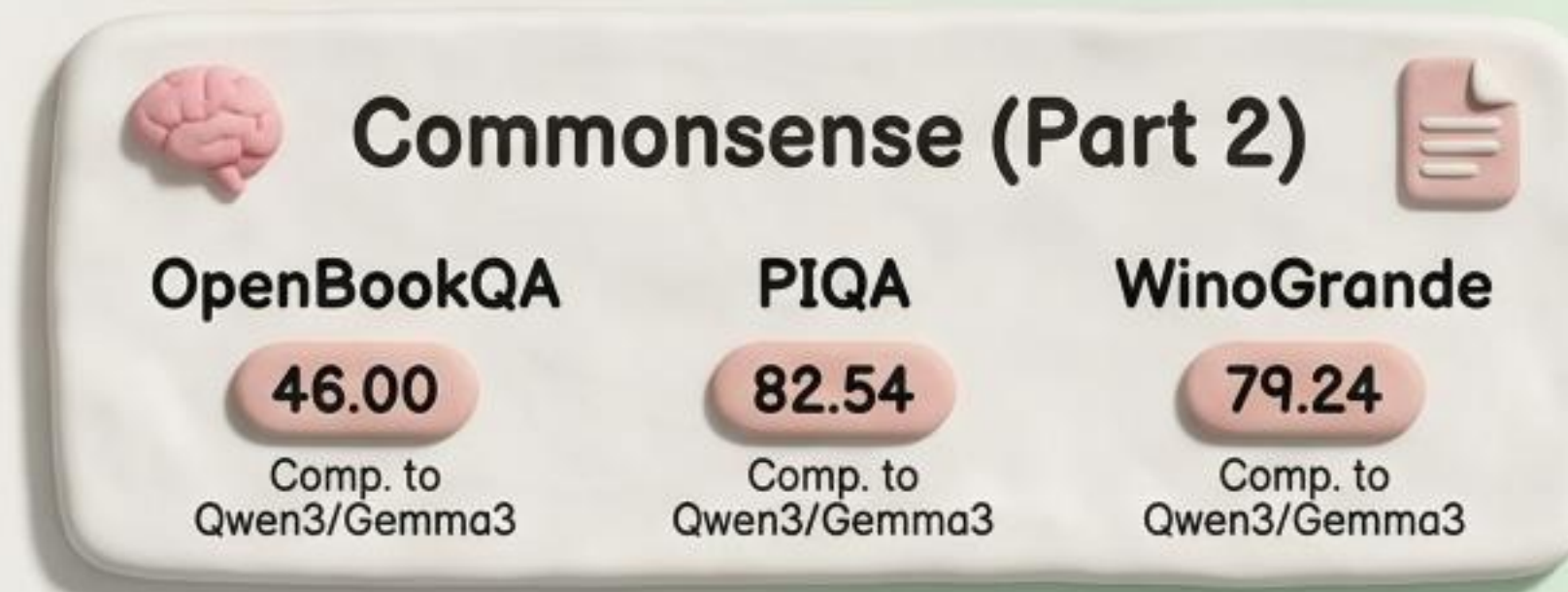
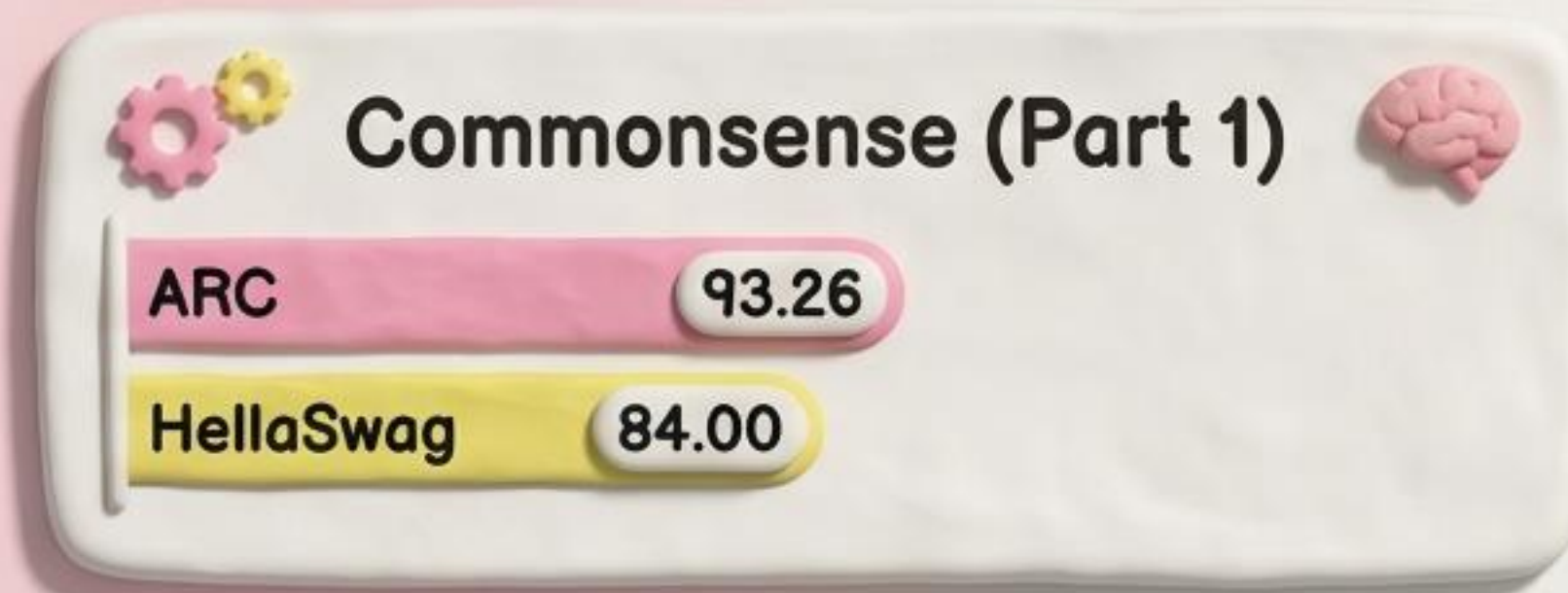
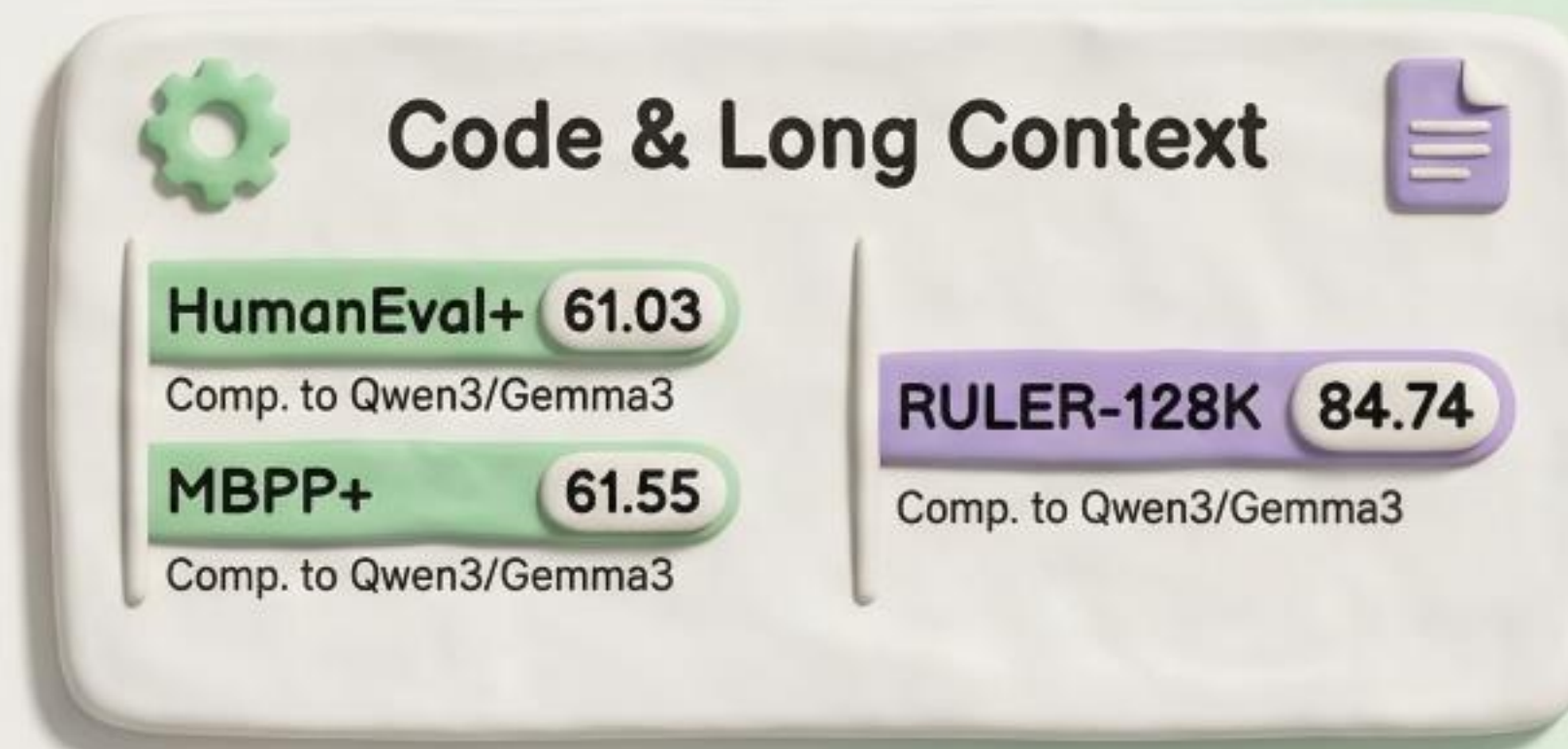
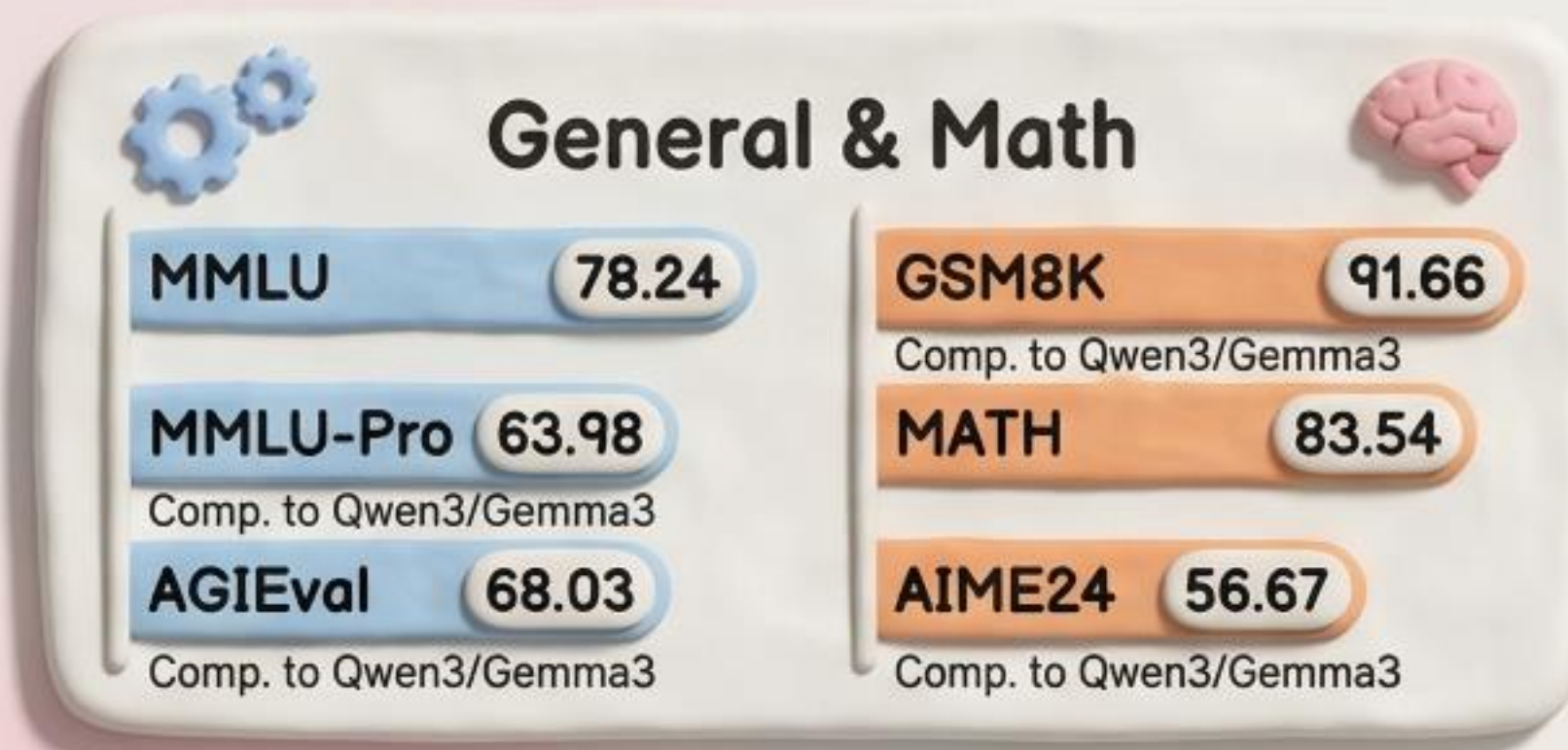
● **Long-Context QA Lateralation**



Long-Context QA

Base Model Evaluation Results

Table 5 Benchmark Comparisons: Competitive Performance



Post-Training Data Distribution

Math
(1.5M Samples)



Math

Coding
(1.1M Samples)



Coding

Science
(2.0M Samples)



Science

Tool-calling
(400K Samples)



Tool-calling

Conversational
(1.5M Samples)



Conversational

Safety
(2K Samples)



Safety

Multilingual
(5.0M Samples)



Multilingual








Total SFT Data

~11.5M Samples

~80B Tokens



Data Generation Methods

-  **Math:** Synthetic, Textbook Extraction
-  **Coding:** Repository Mining, Code Generation
-  **Science:** Research Papers, Database Scraping
-  **Tool-calling:** API Trace Simulation
-  **Conversational:** Dialogue Generation, Crowdsourcing
-  **Safety:** Adversarial Testing, Policy Guidelines
-  **Multilingual:** Translation, Cross-lingual Transfer

Three-Stage SFT Training Process

Stage 1: Foundation



 **10%**
Reasoning-
Stripped

**Full Dataset
&
128k Token
Concatenation**

Stage 2: Skills



**Tool-Calling Focus
(No Concatenation)**

**Full Tool Dataset +
Subsampled Domains**

Stage 3: Refinement



**Long-Context
Reinforcement
(Nemotron-H)**

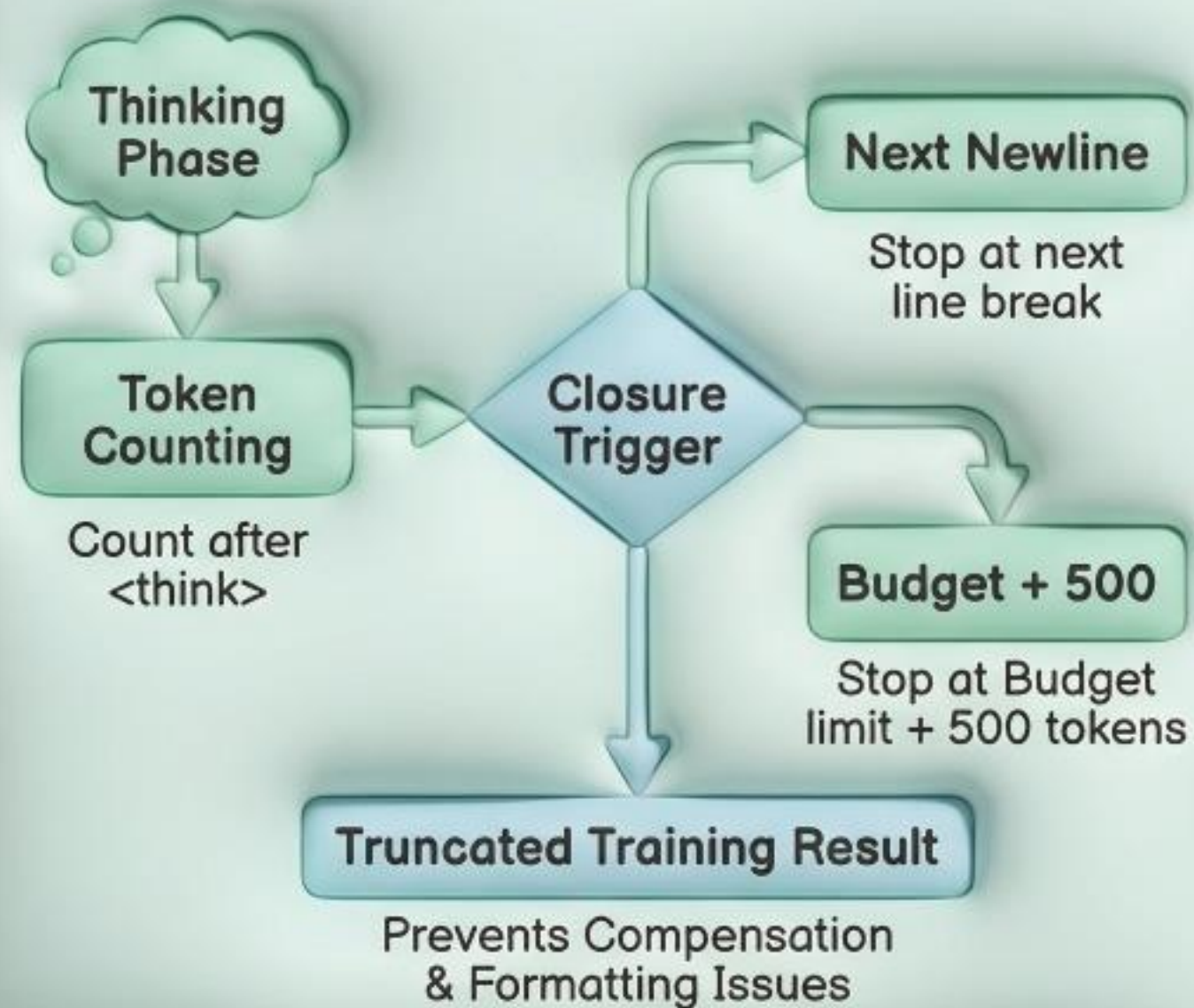
**1-2k Truncated
Traces**

RLHF & Budget Control Mechanism

RLHF Methods & Data Sources



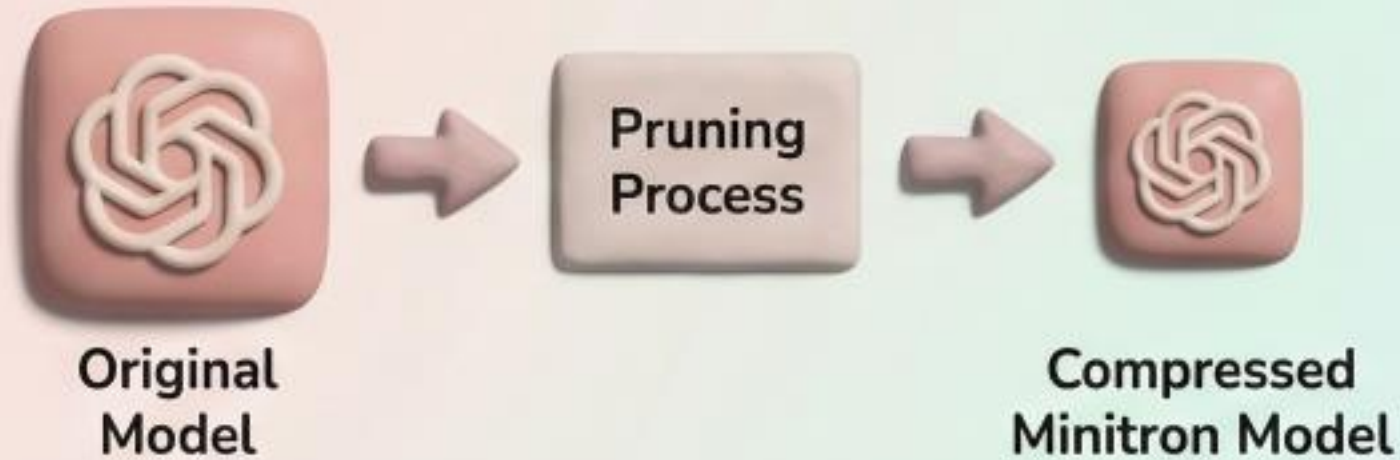
Budget Control Mechanism



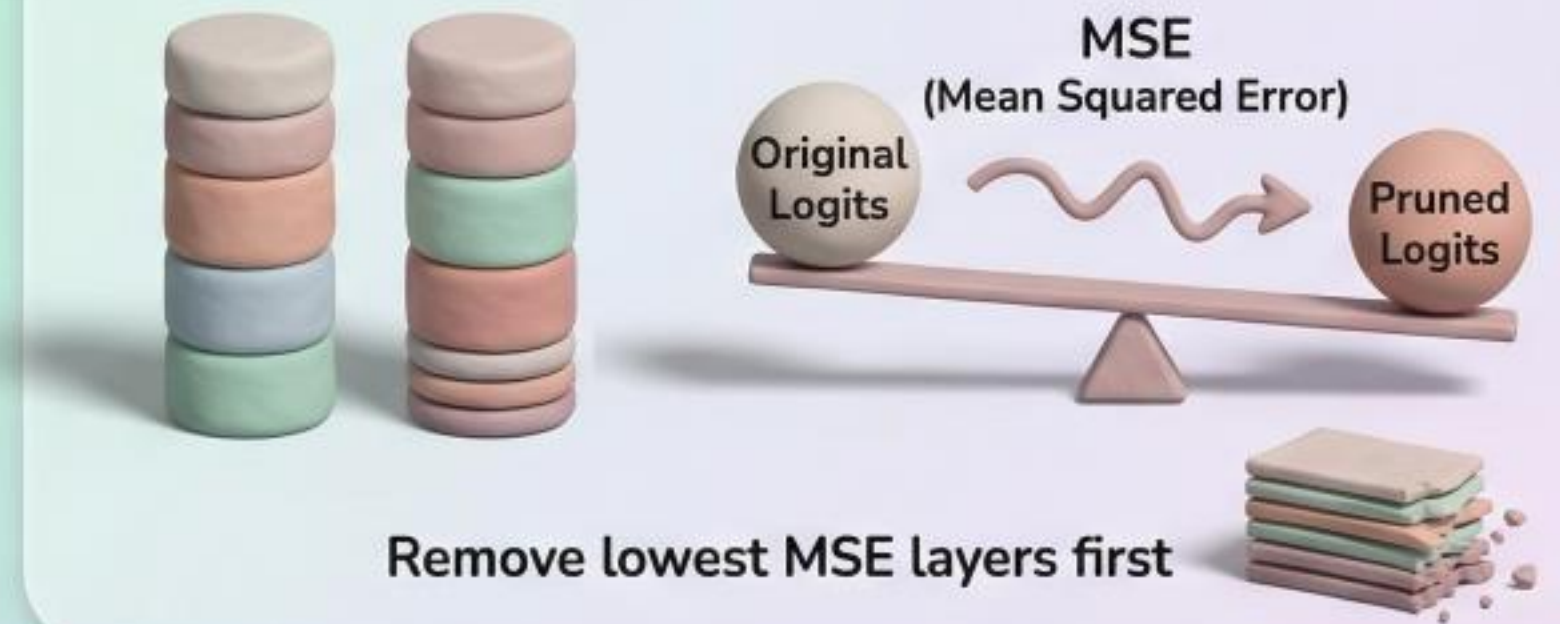
Pruning Strategy: Importance Estimation

Minitron-Based Compression Overview

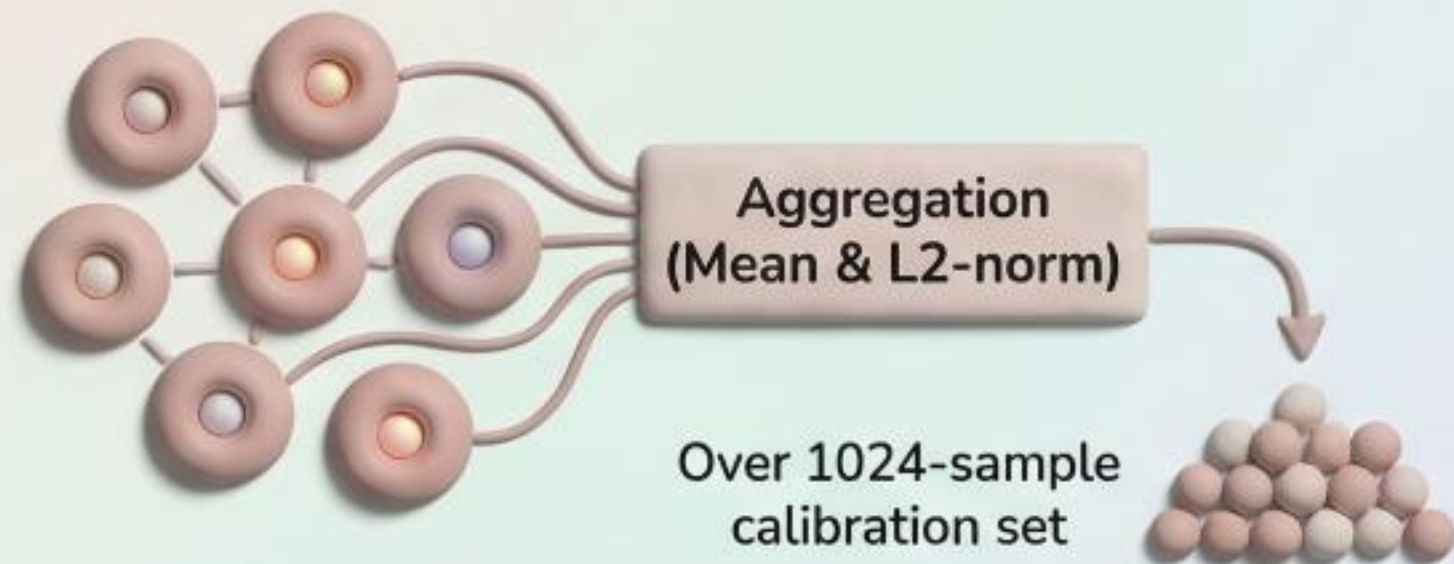
Approach: Iterative pruning based on importance scores



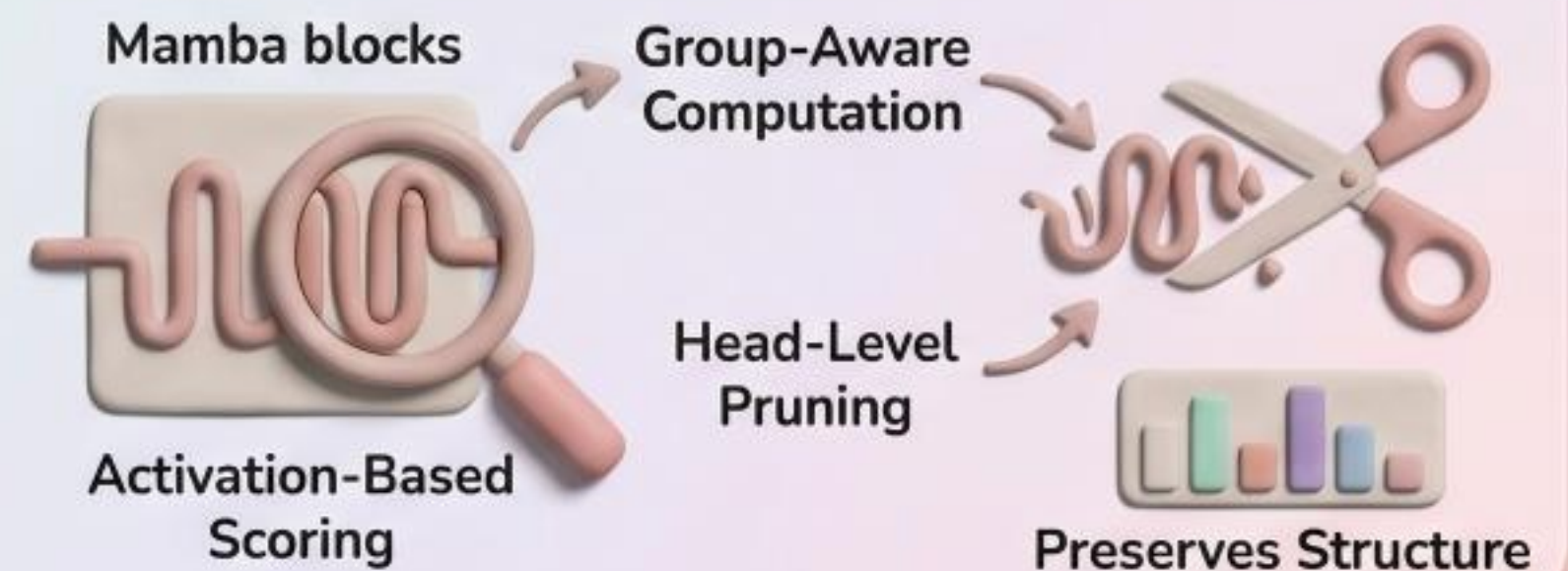
Layer Importance: MSE Computation



FFN/Embedding Importance: Neuron Aggregation



Mamba Importance: Nested Scoring



Architecture Search & Candidate Selection

Identifying the optimal balance of performance and efficiency through depth and width pruning.

Table 10: Top 3 Candidates After Depth+Width Pruning

Candidate 1

**56**
Layers

**4480**
Hidden Dim

**17920**
FFN


**112**
Mamba
Heads


**8.92B**
Params


**59.07%**
Accuracy


**161.02**
Throughput


Candidate 2


**56**
Layers


**4480**
Hidden Dim

**15680**
FFN

**128**
Mamba
Heads

**8.89B**
Params

**63.02%**
Accuracy

**156.42**
Throughput

Selected for Best Accuracy-Speed Balance

Candidate 3

**56**
Layers

**4800**
Hidden Dim

**14400**
FFN

**120**
Mamba
Heads

**8.97B**
Params

**62.94%**
Accuracy

**155.86**
Throughput

Candidate 2 selected for best accuracy-speed balance.

Distillation Pipeline Stages

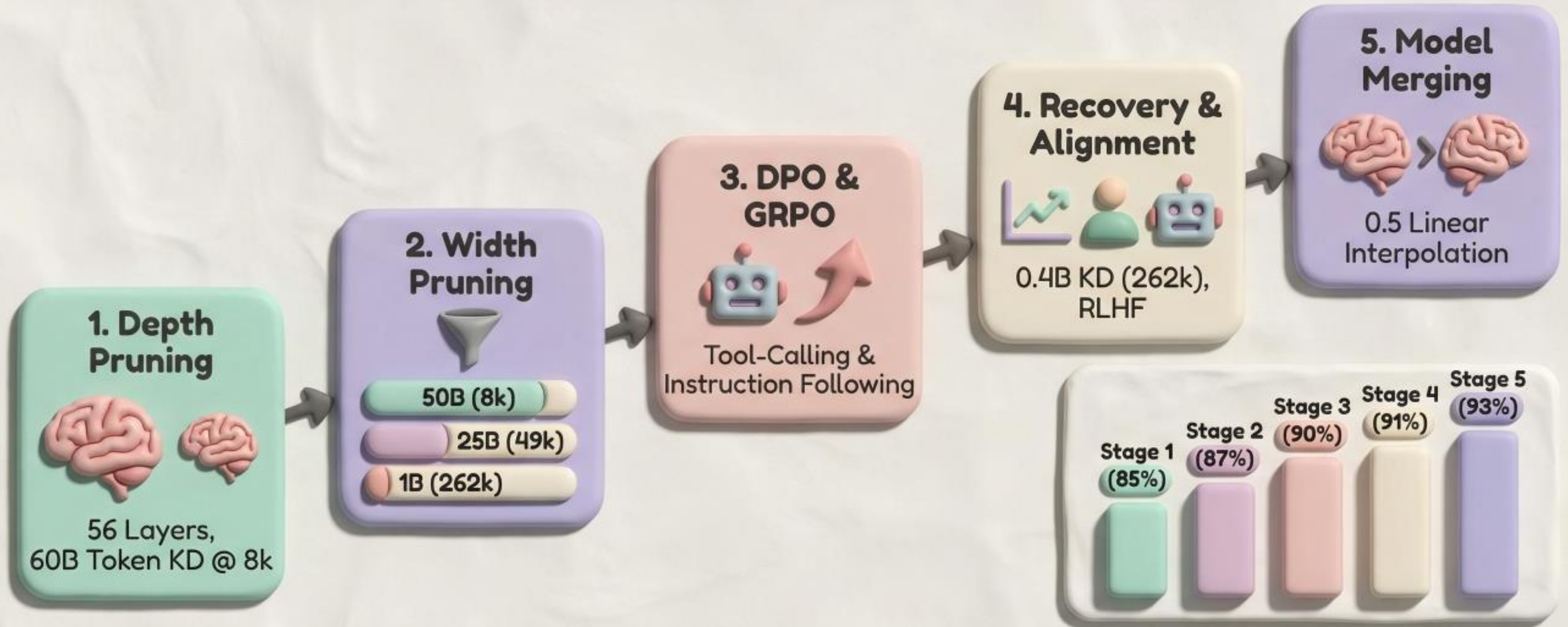
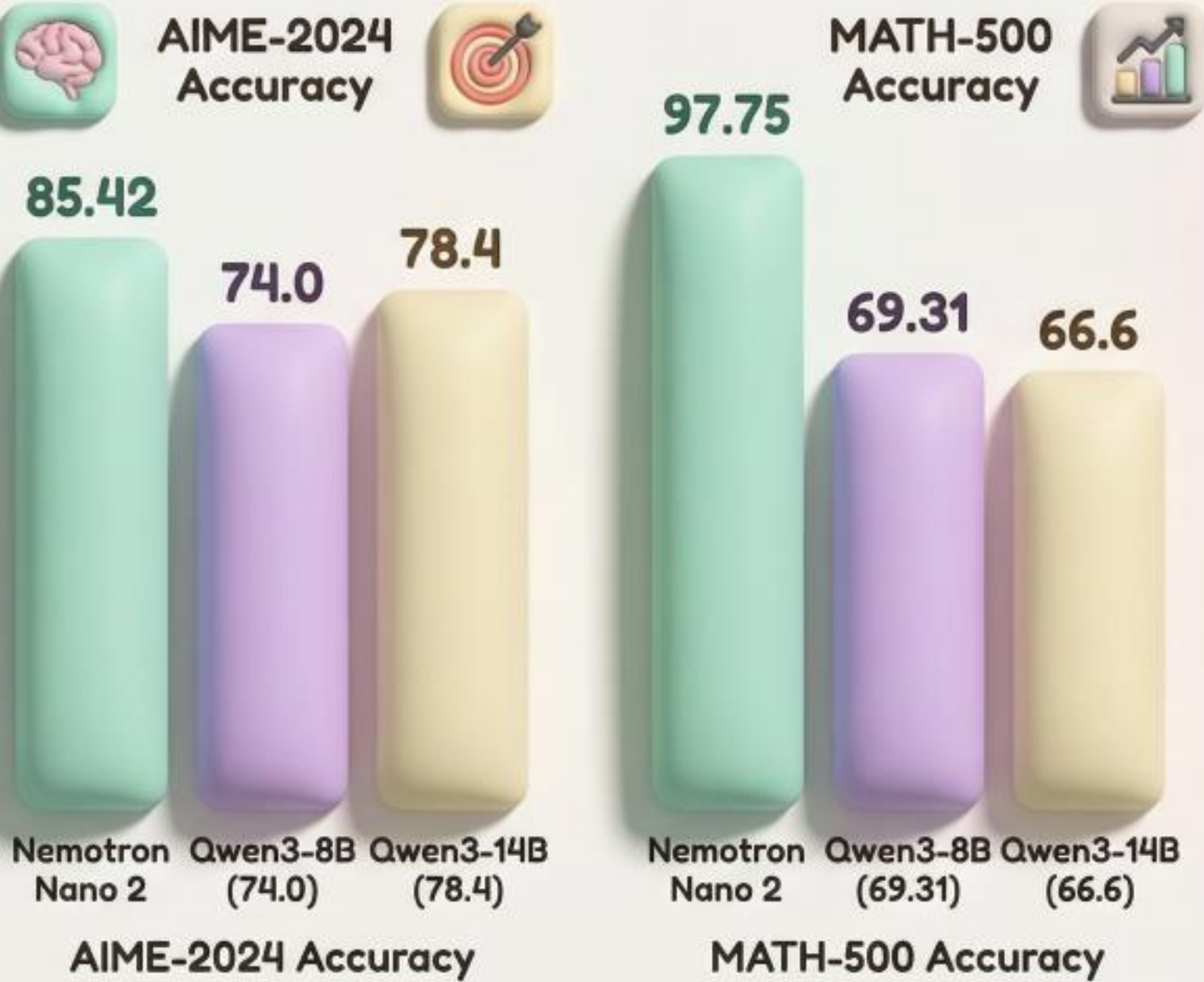


Figure 6: Accuracy Progression

Final Model Performance

Table 8 Evaluation Results (Reasoning ON)






Benchmark	Nemotron Nano 2	Qwen3-8B	Qwen3-14B
AIME-2024	85.42	74.0	78.4
AIME-2025	76.25	75.83	81.53
MATH-500	97.75	69.31	66.6
GPQA-Diamond	64.48	59.61	64.53
LiveCodeBench	70.79	59.5	63.08
BFCL v3	89.81	89.39	91.32
RULER@128k	66.98	66.34	68.01
Arena Hard	83.36	74.13	73.55

Released Models & Datasets







Hugging Face Models

-  **Nemotron-Nano-9B-v2**
↳ (Aligned & Pruned)
-  **Nemotron-Nano-9B-v2-Base**
↳ (Pruned Base)
-  **Nemotron-Nano-12B-v2-Base**
↳ (Original Base)



Datasets

6T+ Pre-Training Datasets (6T+ Tokens)

-  Nemotron-**CC-v2**
-  Nemotron-**CC-Math-v1**
-  Nemotron-**Pretraining-Code-v1**
-  Nemotron-**Pretraining-SFT-v1**

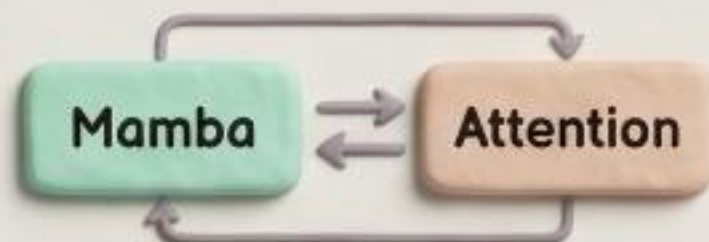


Post-Training Dataset

- Nemotron-**Post-Training-Dataset-v2**
(5 Language Extensions)

Key Technical Insights

Hybrid Mamba-Transformer Architecture



Enables 3-6x throughput with competitive accuracy

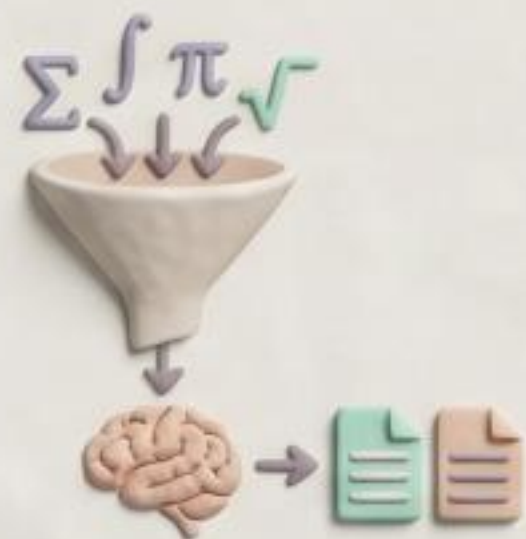


3-6x
Throughput



Competitive Accuracy

Math-Specific Extraction Pipeline



Superior mathematical reasoning data; Three-phase curriculum training



Phase 1:
Diversity



Phase 2:
Quality



Phase 3:
Curriculum

Reasoning SFT & 512k Context



Improved
MMLU-Pro



512k
Context

Fundamental Reasoning SFT significantly improves MMLU-Pro; 512k context reduces document fragmentation

Budget Control & Minitron Compression



Budget
Control
Training



A10G
Inference

Prevents compensation behaviors; Maintains accuracy, enables A10G inference