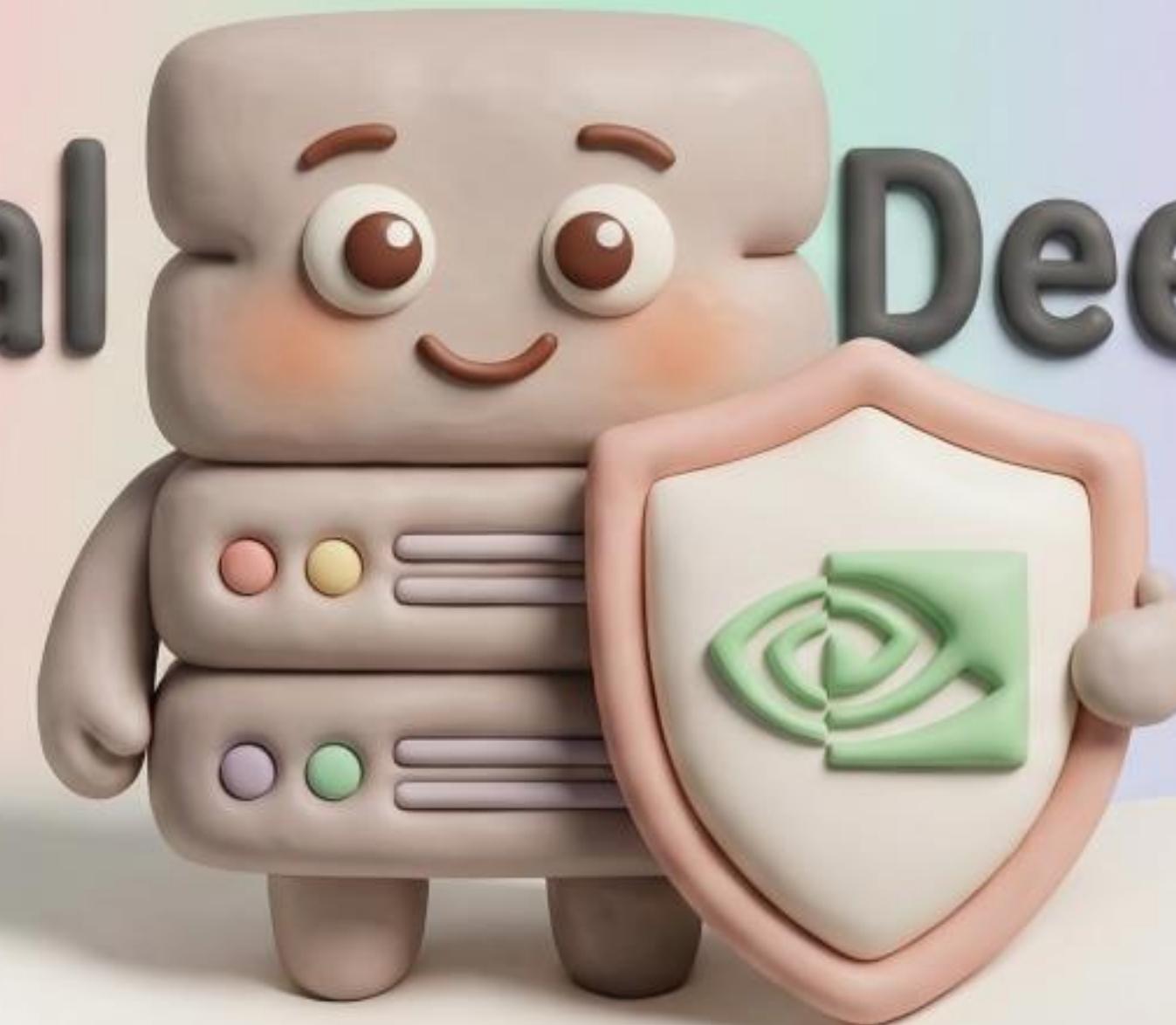


# NVIDIA Nemotron Nano 2: Technical Deep Dive



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

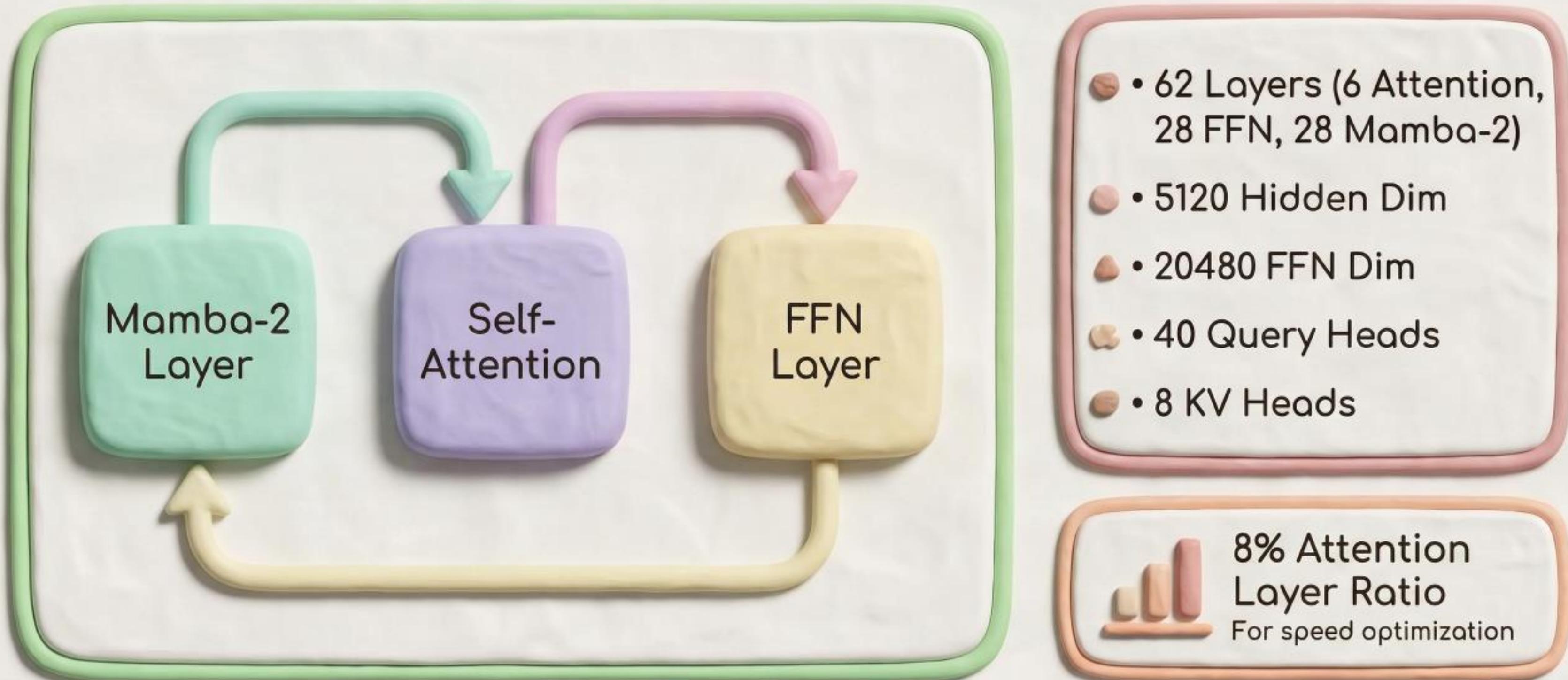
- 9B Parameters
- 128k Context Length
- 20T Token Training



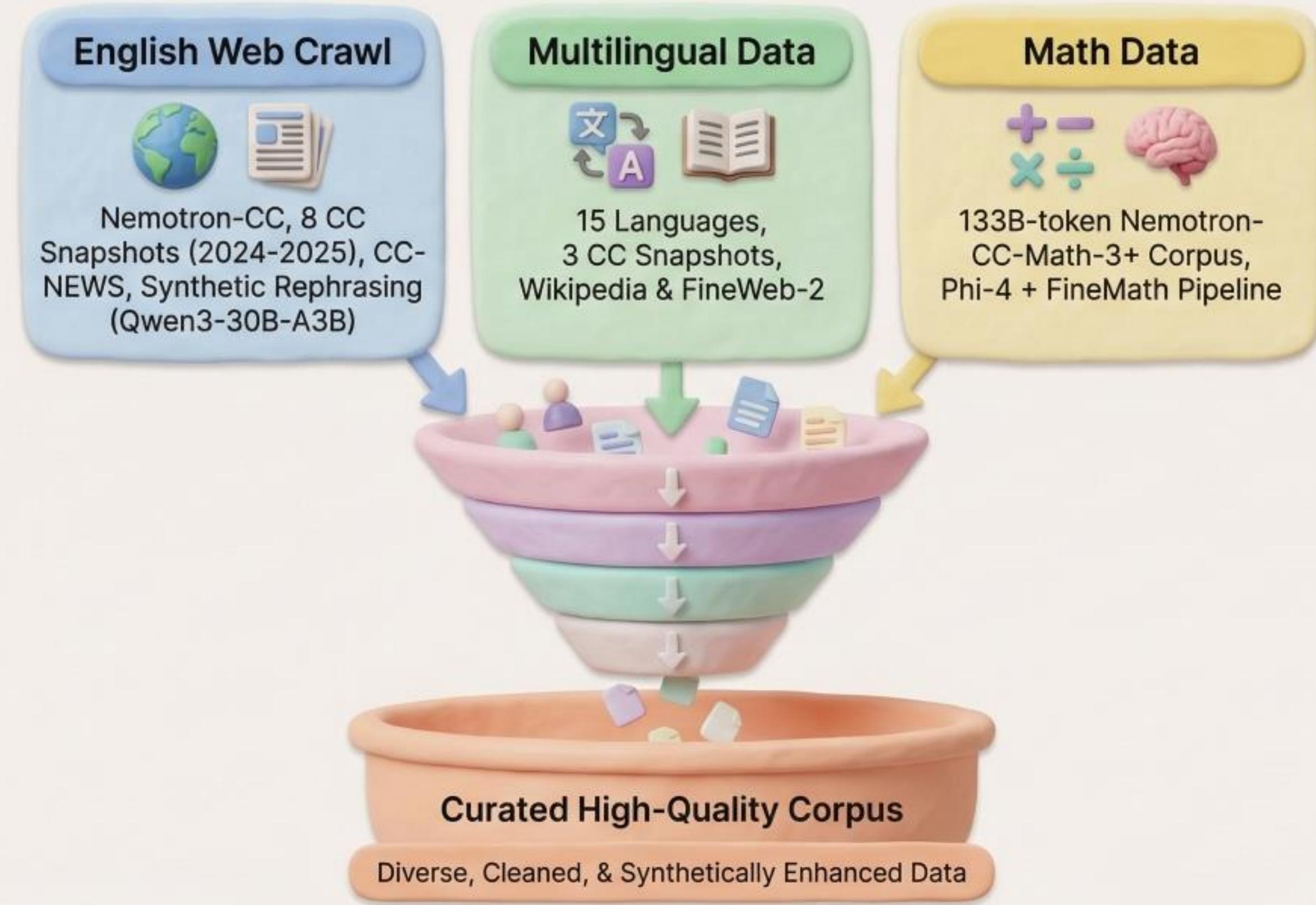
## Performance

- Competitive Accuracy
- Dramatically Higher Inference Speed

# Model Architecture: Hybrid Design



# Pre-Training Data: Curated Sources



# Pre-Training Data: Math Extraction Pipeline

## Overview: Specialized Math Extraction Process



Aggregate Math URLs  
from Prior Datasets



Refetch HTML from  
98 CC snapshots  
(2014-2024)



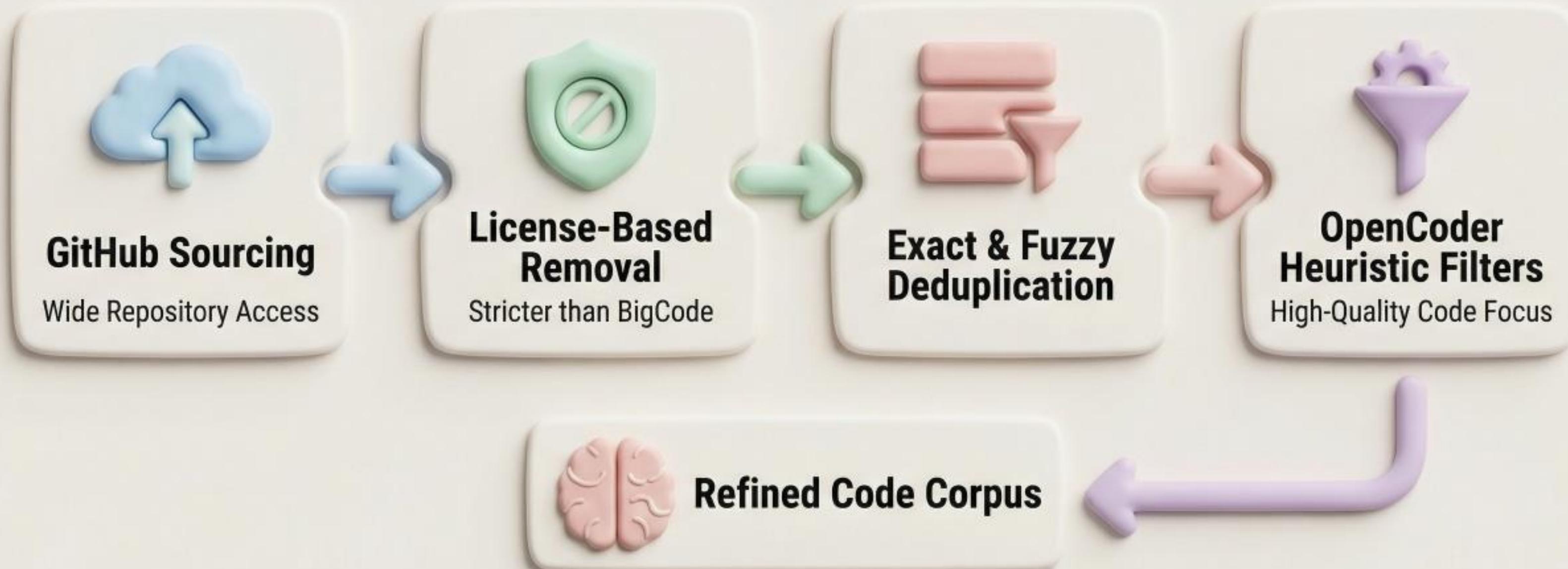
Render with lynx  
browser, apply Phi-4 for  
boilerplate removal &  
LaTeX standardization



FineMath classifier  
for quality filtering,  
MinHash LSH  
deduplication

# 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



Enhanced ability to discern valid reasoning from plausible but incorrect choices.



# Training Configuration & FP8 Recipe

## Training Hyperparameters



20T Tokens Horizon

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



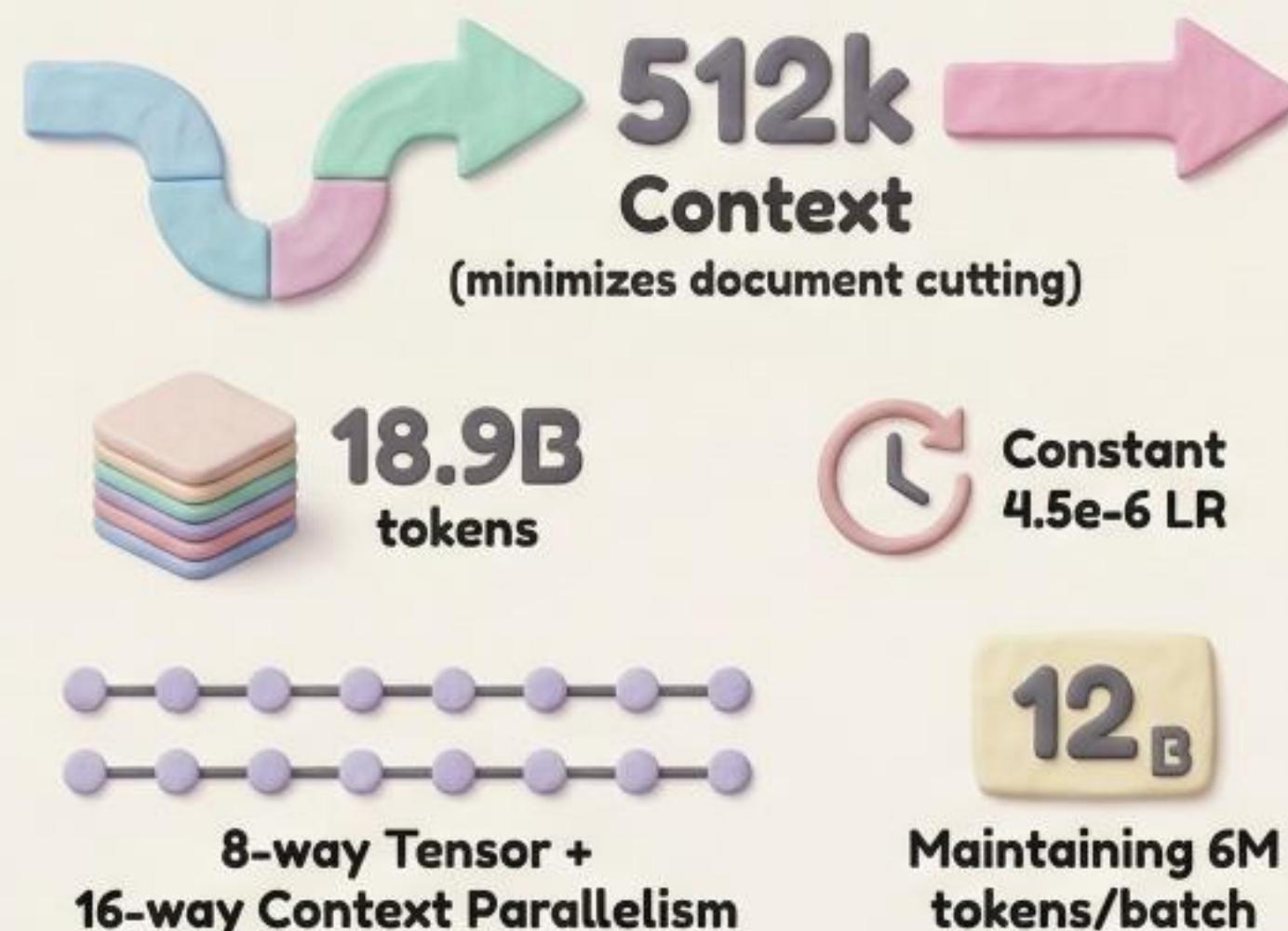
FP8 Parameter  
All-Gather



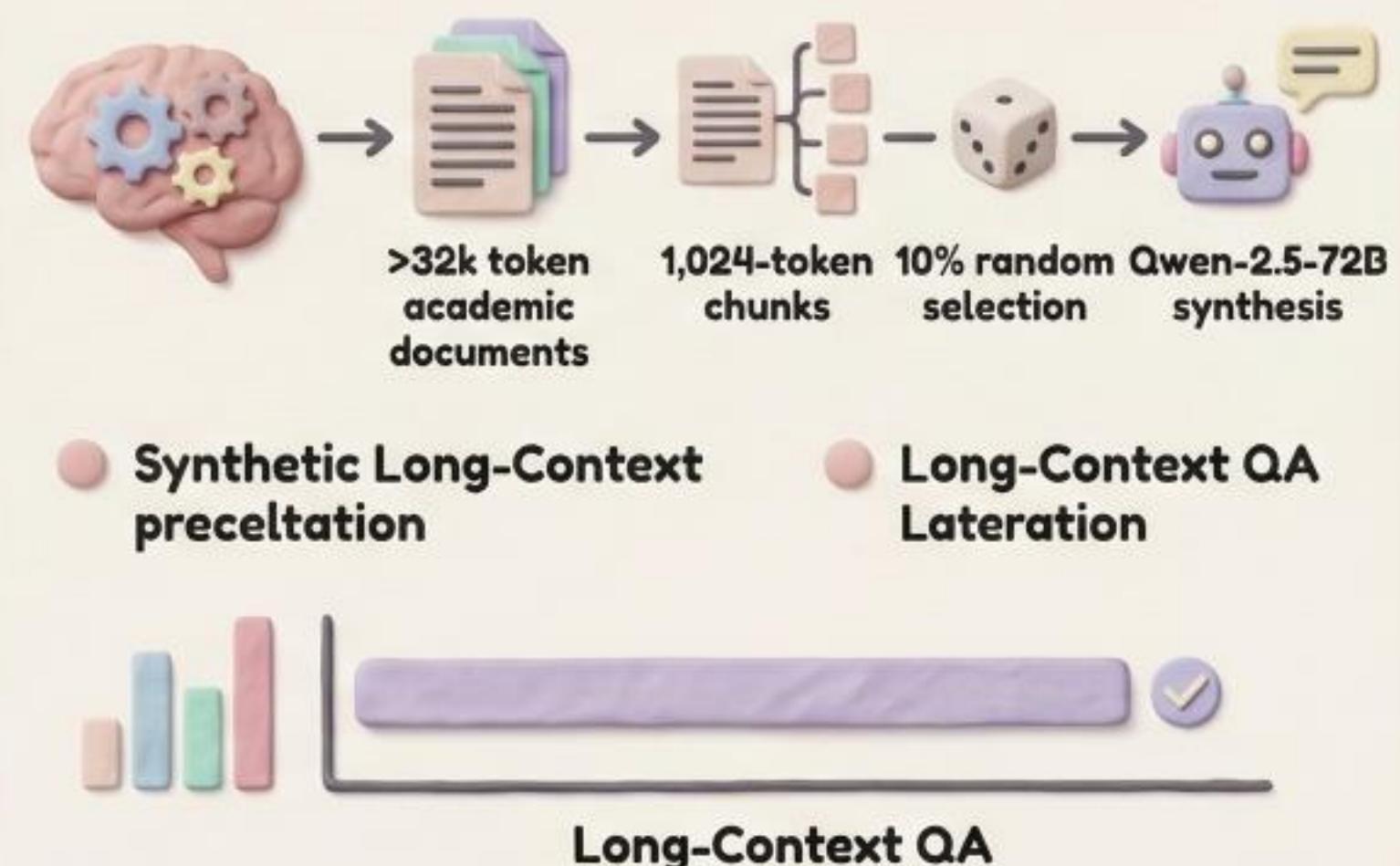
FP32  
Master  
Weights

# Long-Context Extension Strategy

## Phase LC Continuous Pretraining

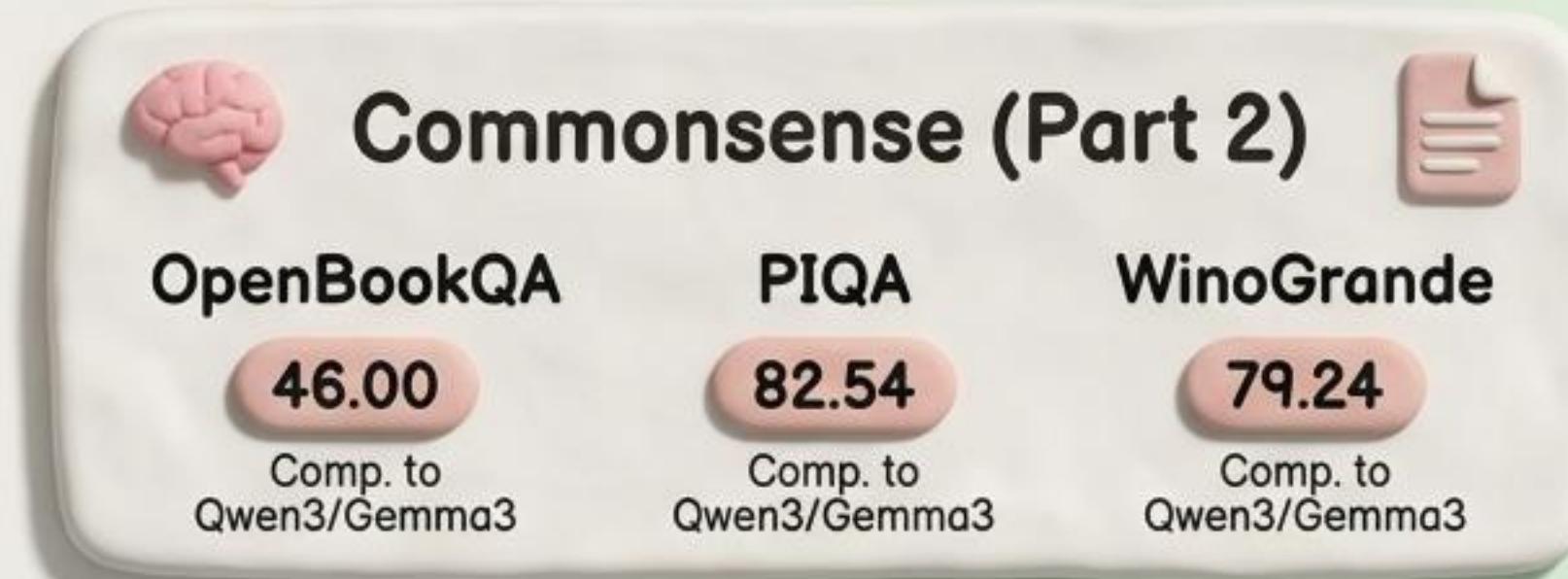
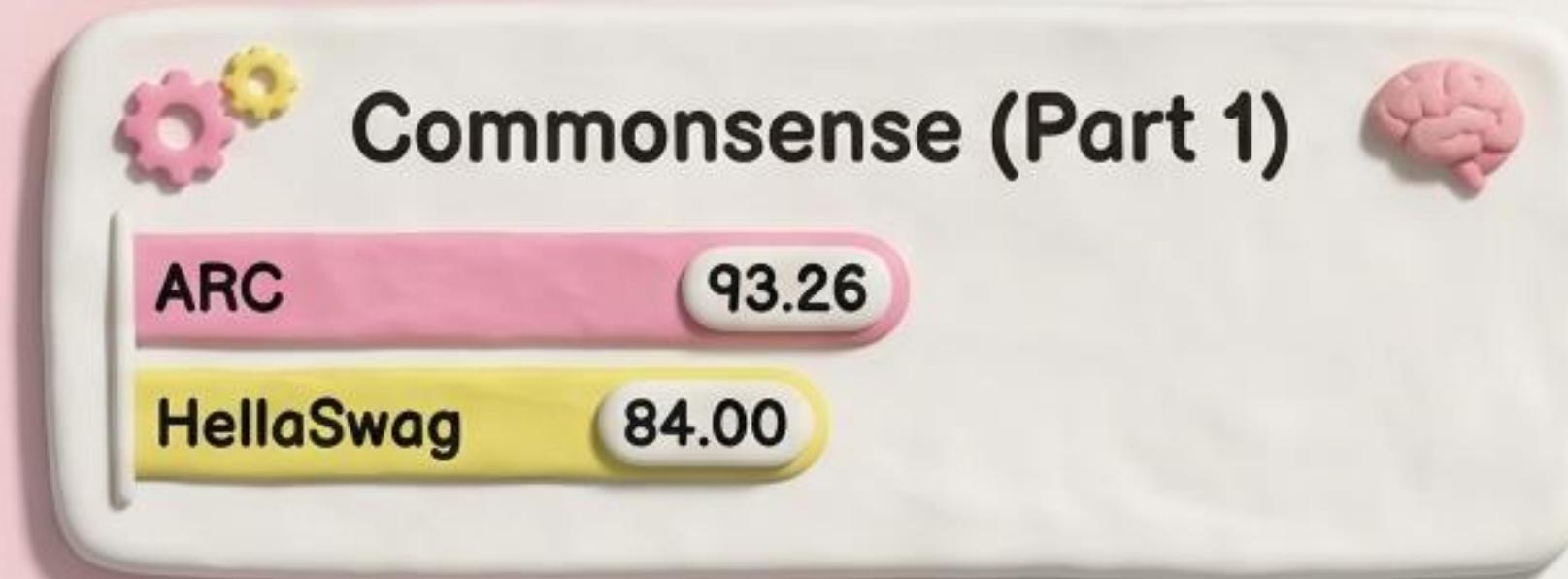
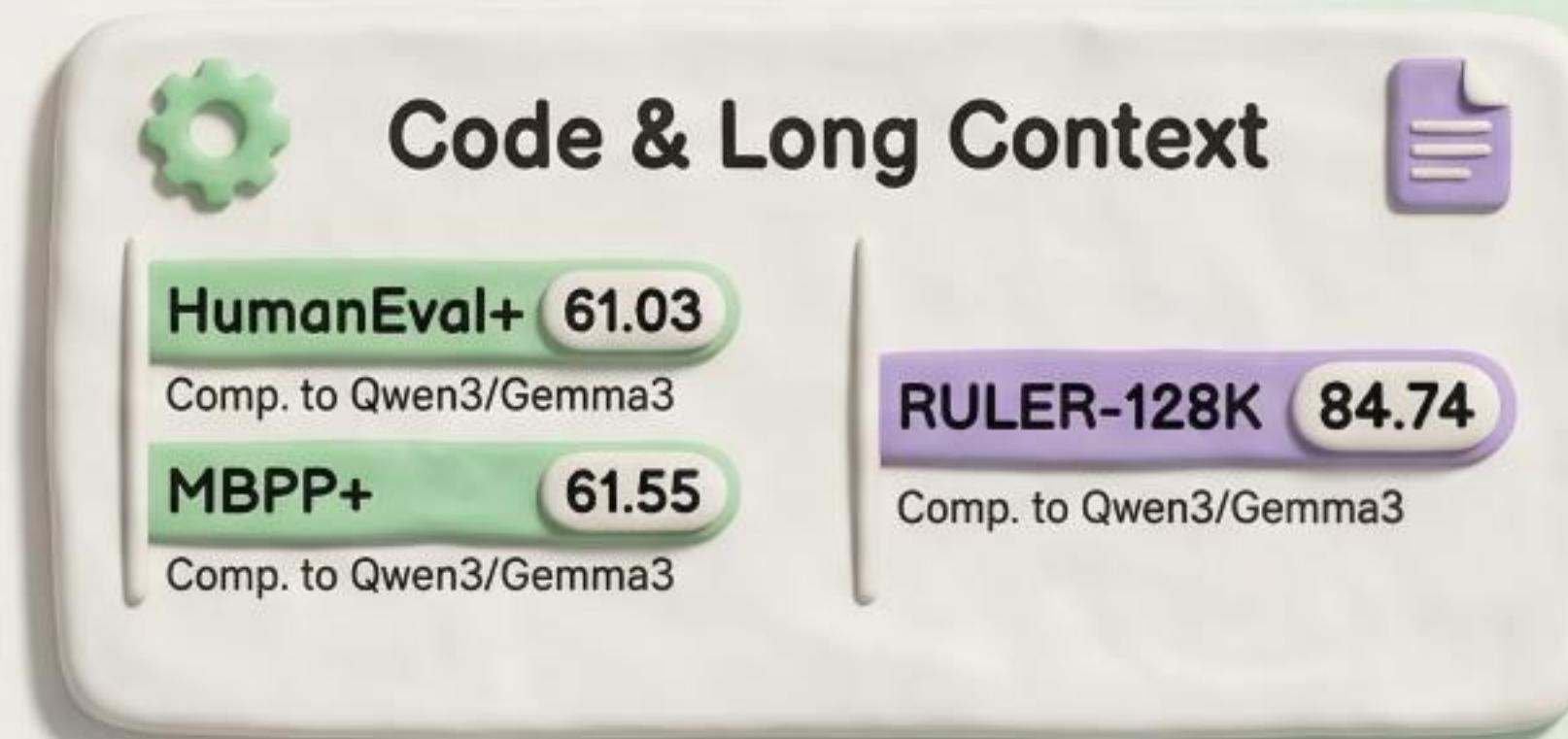
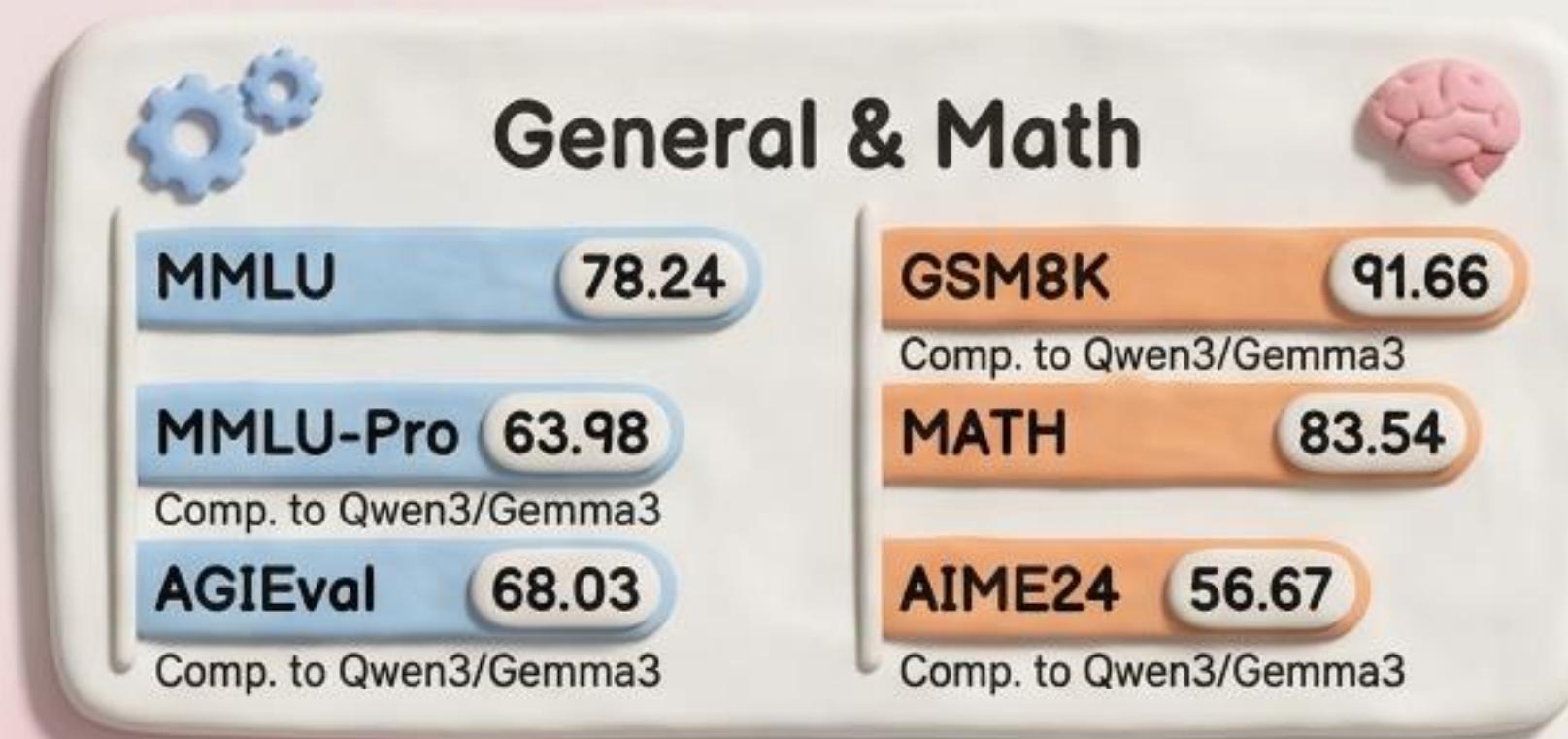


## Synthetic Long-Context QA Generation



# Base Model Evaluation Results

Table 5 Benchmark Comparisons: Competitive Performance



# Post-Training Data Distribution

Multilingual  
(5.0M Samples)



## 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



# RLHF & Budget Control Mechanism

## RLHF Methods & Data Sources



**IFEval RL**  
16K Prompts,  
Rule-based Verifier

**DPO for Tool-calling**  
WorkBench, Multi-step



**GRPO for Helpfulness**  
HelpSteer3 Contexts,  
Qwen-based Reward



## Budget Control Mechanism



**Token Counting**

Count after  
<think>

**Next Newline**

Stop at next  
line break

**Closure Trigger**

**Budget + 500**

Stop at Budget  
limit + 500 tokens

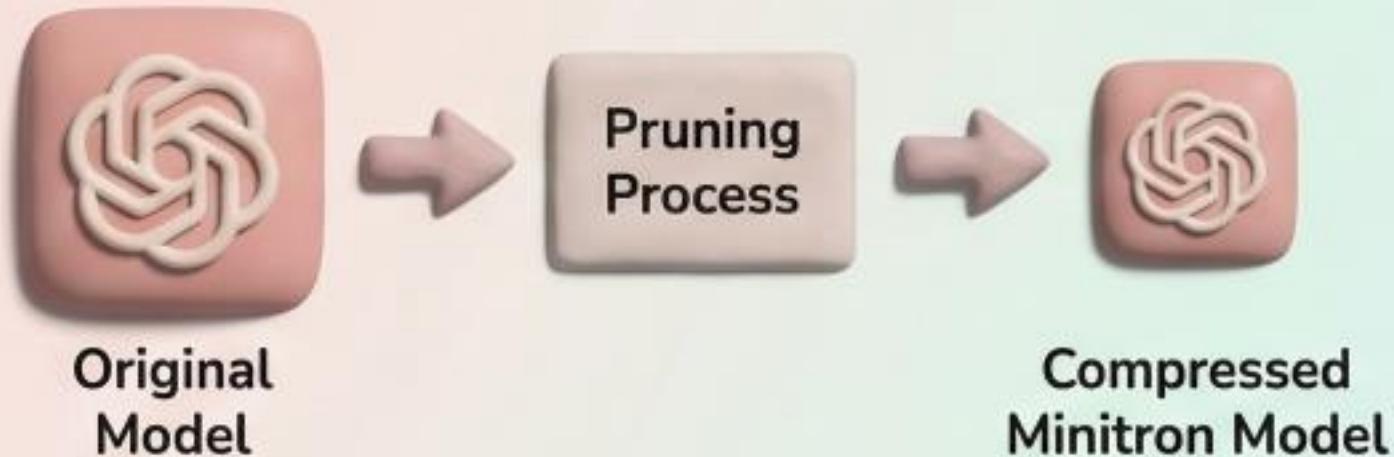
**Truncated Training Result**

Prevents Compensation  
& Formatting Issues

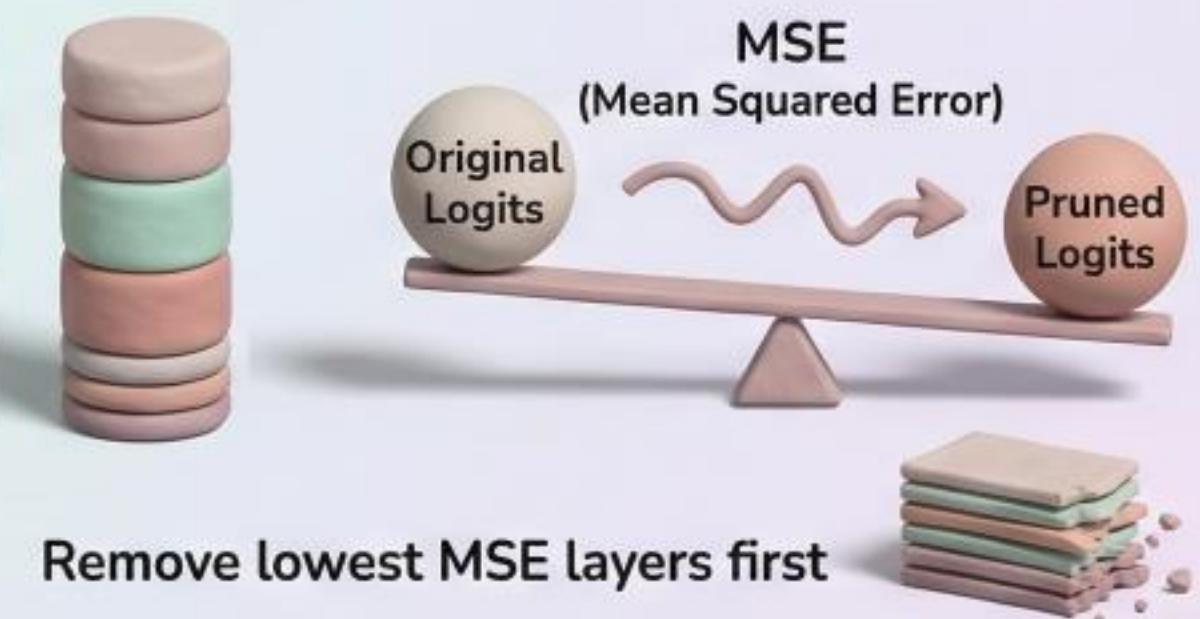
# Pruning Strategy: Importance Estimation

## Minitron-Based Compression Overview

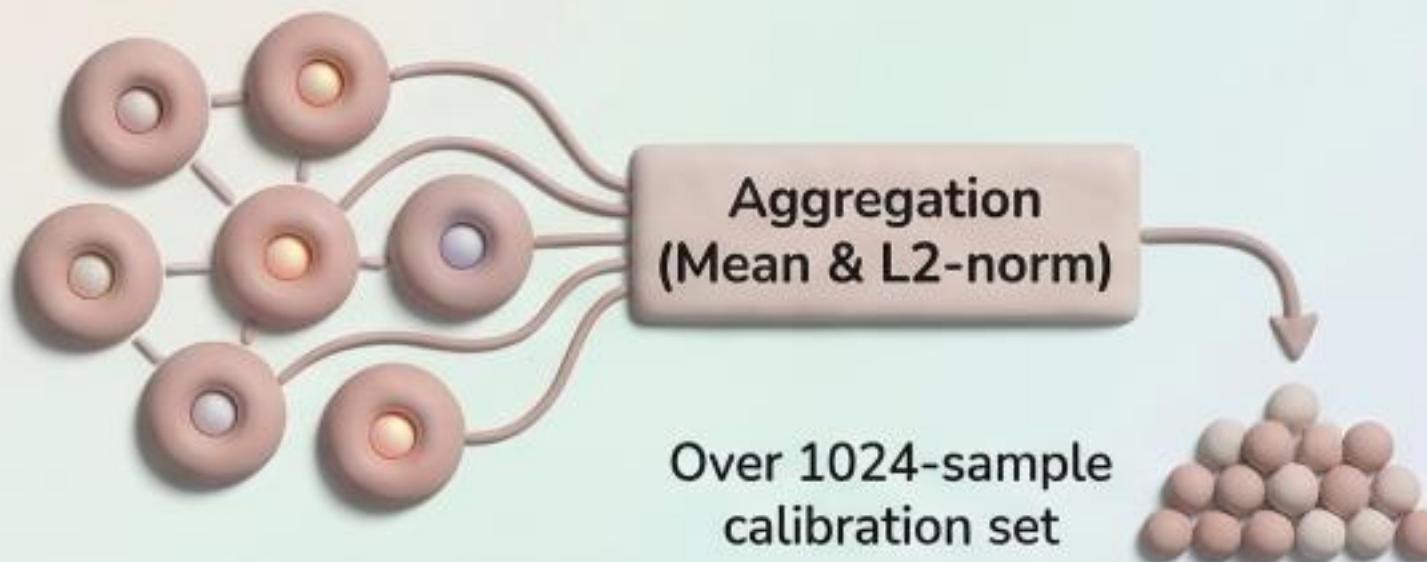
Approach: Iterative pruning based on importance scores



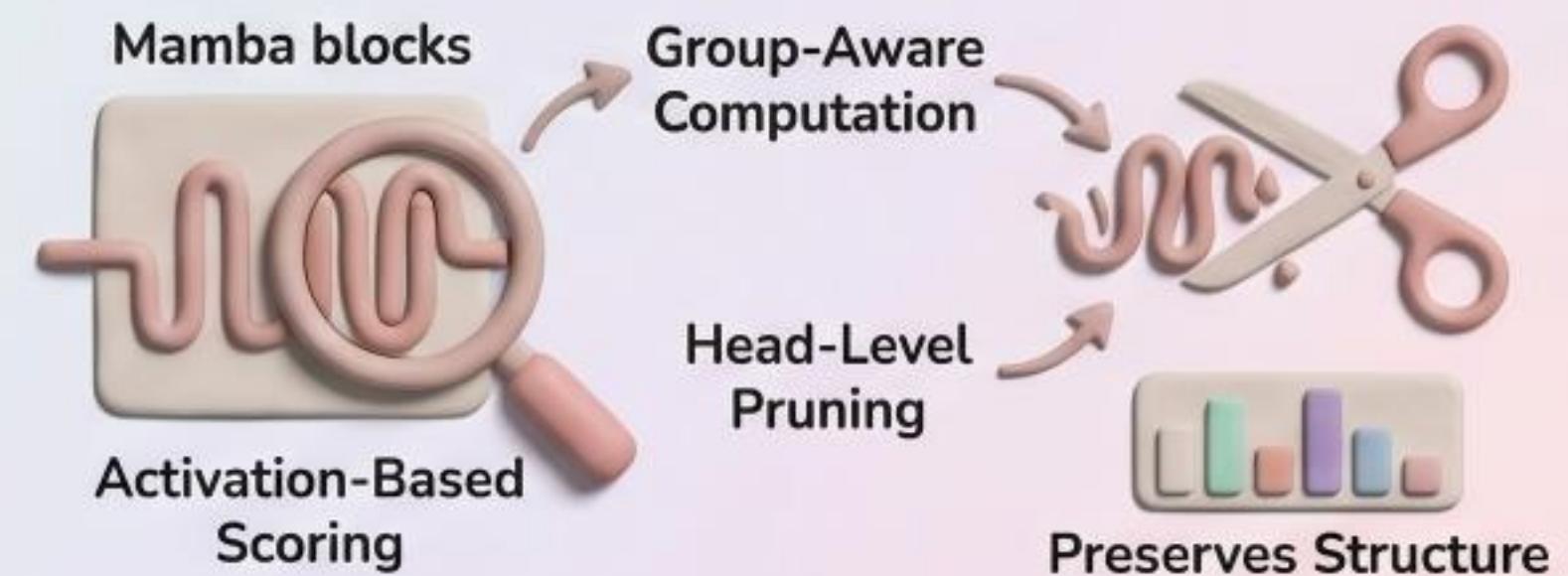
## Layer Importance: MSE Computation



## 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

<b>56</b> Layers	<b>4480</b> Hidden Dim	<b>17920</b> FFN	<b>112</b> Mamba Heads

<b>8.92B</b> Params	<b>59.07%</b> Accuracy	<b>161.02</b> Throughput

## Candidate 2

<b>56</b> Layers	<b>4480</b> Hidden Dim	<b>15680</b> FFN	<b>128</b> Mamba
Selected for Best Accuracy-Speed Balance			

<b>8.89B</b> Params	<b>63.02%</b> Accuracy	<b>156.42</b> Throughput

## Candidate 3

<b>56</b> Layers	<b>4800</b> Hidden Dim	<b>14400</b> FFN	<b>120</b> Mamba Heads

<b>8.97B</b> Params	<b>62.94%</b> Accuracy	<b>155.86</b> Throughput

Candidate 2 selected for best accuracy-speed balance.

# Distillation Pipeline Stages

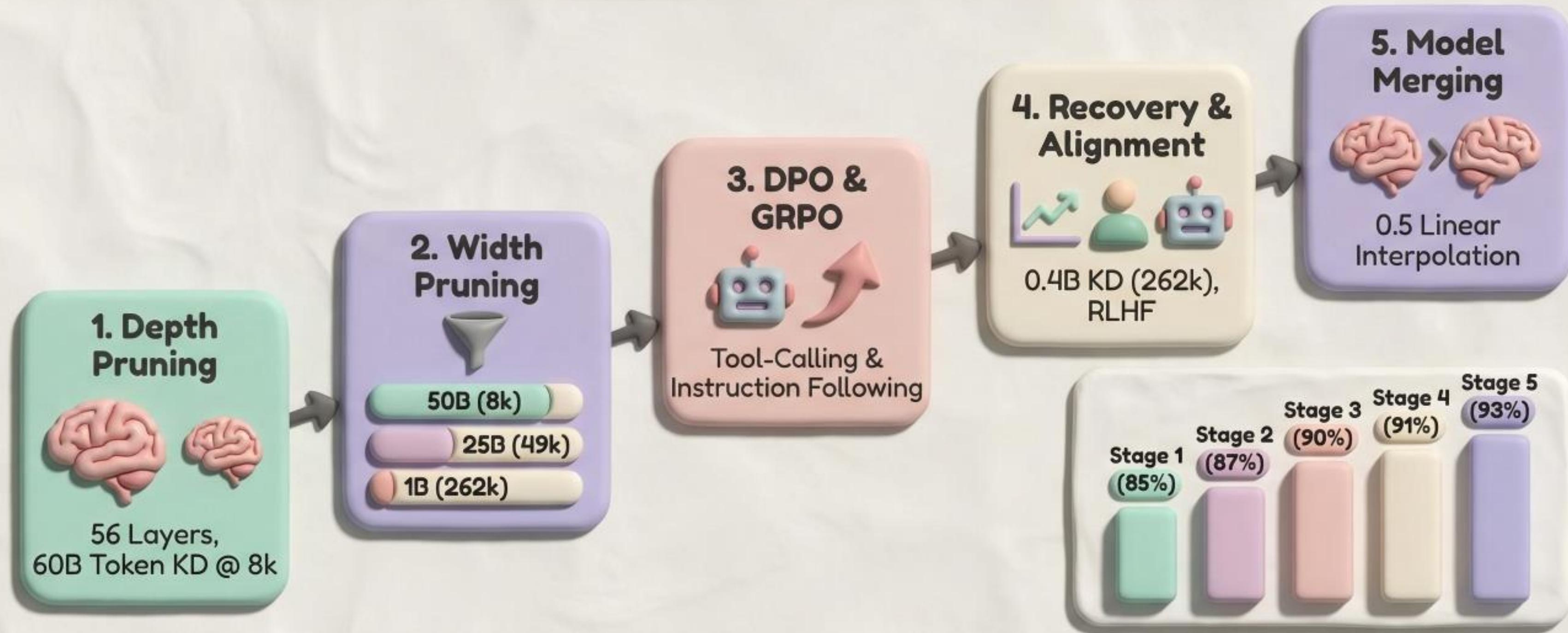


Figure 6: Accuracy Progression

# Final Model Performance

Table 8 Evaluation Results (Reasoning ON)



85.42

74.0

78.4

Nemotron Nano 2 (74.0)

Qwen3-8B (78.4)

97.75

MATH-500 Accuracy



69.31 66.6

Nemotron Nano 2 (69.31)  
Qwen3-8B (66.6)  
Qwen3-14B

AIME-2024 Accuracy

MATH-500 Accuracy

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

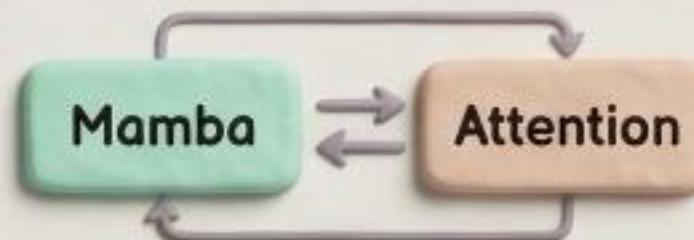


### Post-Training Dataset

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

# Key Technical Insights

## Hybrid Mamba-Transformer Architecture



Enables 3-6x throughput  
with competitive accuracy

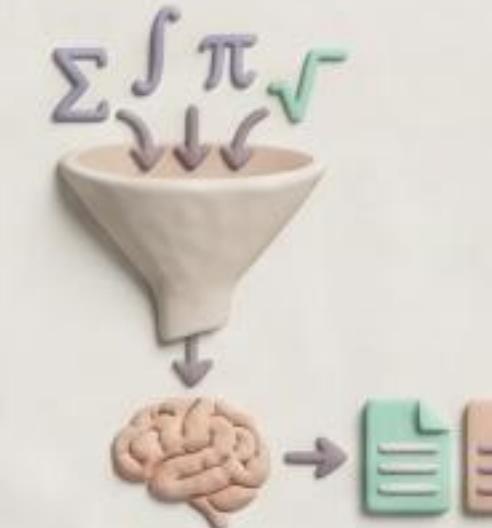


**3-6x**  
Throughput



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