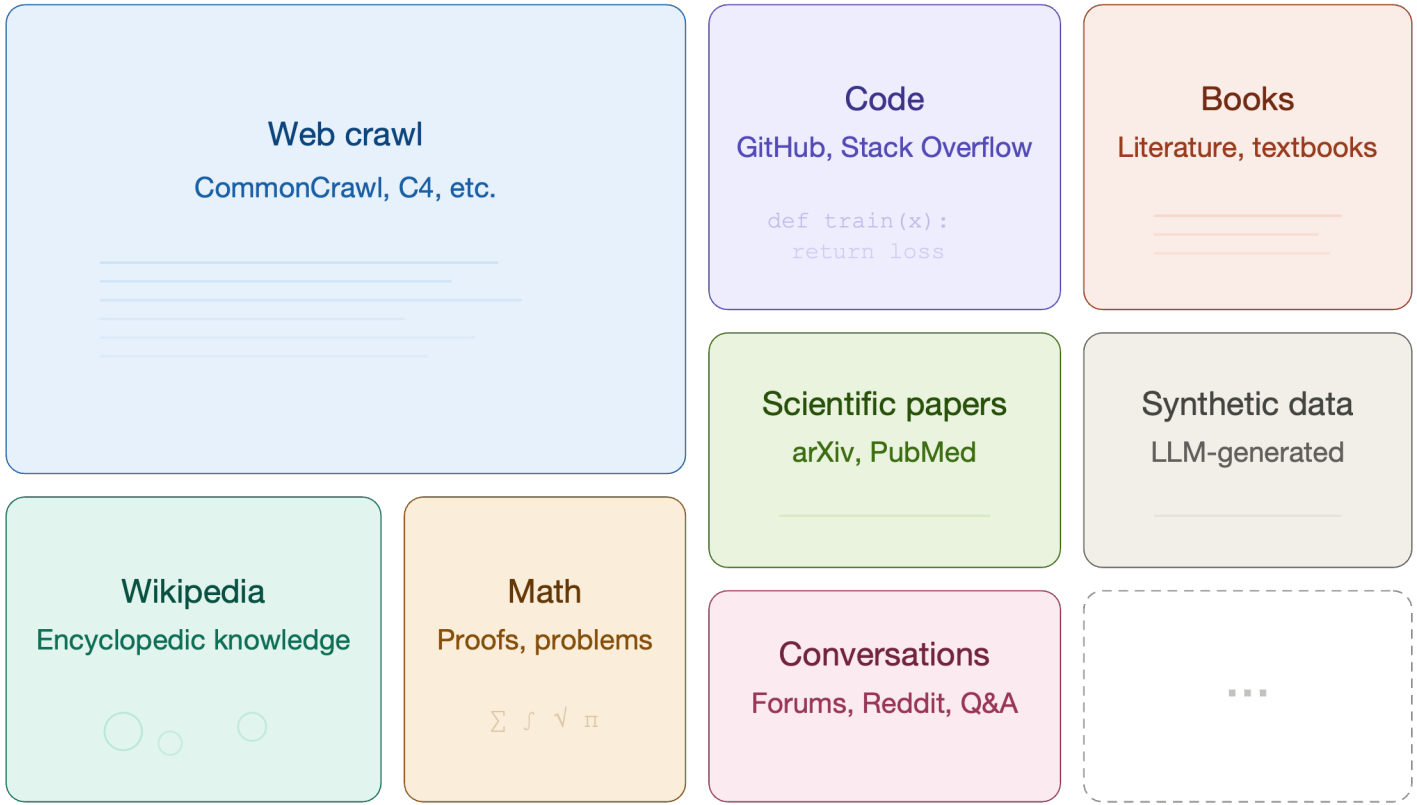


Institute of Foundation Models
Mohamed bin Zayed University of Artificial Intelligence

Diversity First: Lessons from Shaping LLM Training Data at Scale

Mikhail Yurochkin

Data Composition



Data Mixing

$$\text{Data: } D = \{d_1, \ldots, d_N\}$$

$$\text{Training: } \arg\min_{\theta} \mathbb{E}_{d \sim p_\tau(D)} \ell(\text{LLM}_\theta, d)$$

$$p_\tau(D) = \sum_n w_\tau(d_n)\delta_{d_n}$$

$$\text{Data mixing: } \arg\max_{\tau} \text{Evals}(\text{LLM}_\theta)$$



$$\text{such that } \theta = \arg\min_{\theta} \mathbb{E}_{d \sim p_\tau(D)} \ell(\text{LLM}_\theta, d)$$

Does MMLU or HellaSwag of base model correlate with agentic/reasoning/etc ability of the final model?

	Qwen2.5-72b	Llama3.1-70b	Llama3.0-70b
HellaSwag	87.6	85.0	**88.0**
Winogrande	83.9	79.8	**85.3**
PIQA	83.5	84.3	**84.6**

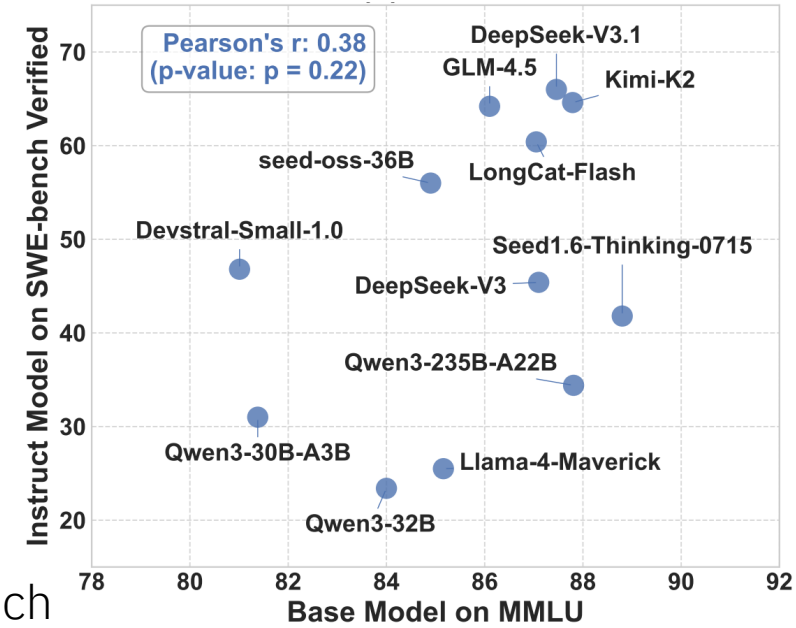


Image from APTBench

Convert post-training Evals to MCQ

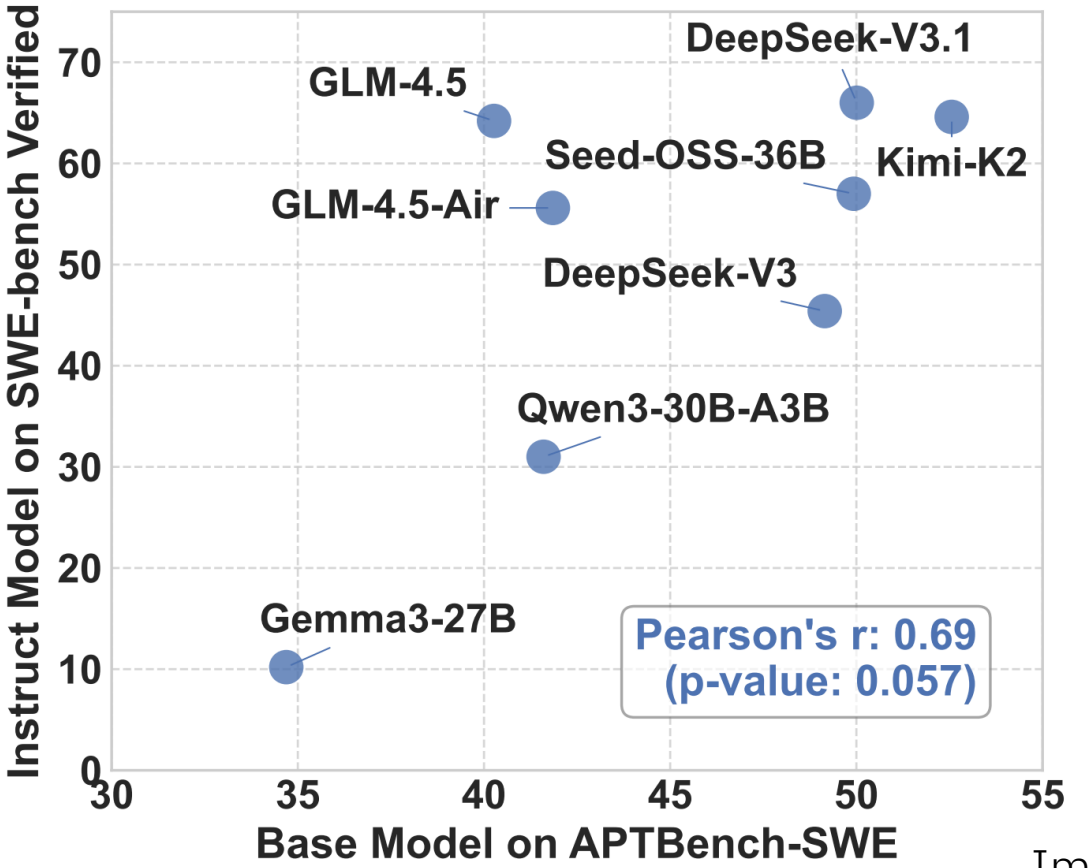


Image from APTBench

Aligning with post-training

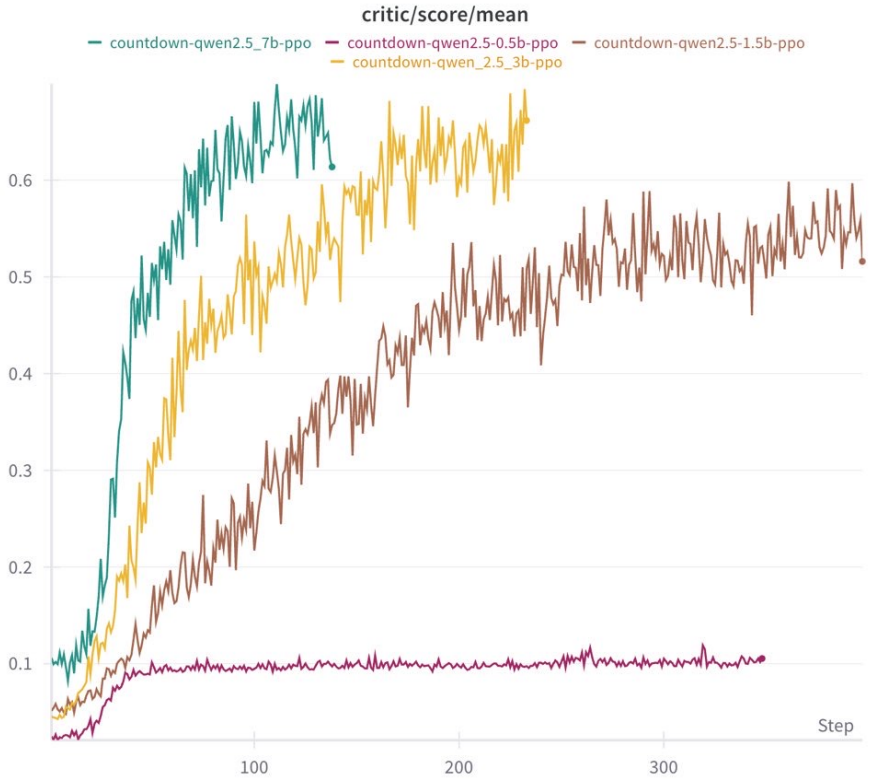


Image from TinyZero

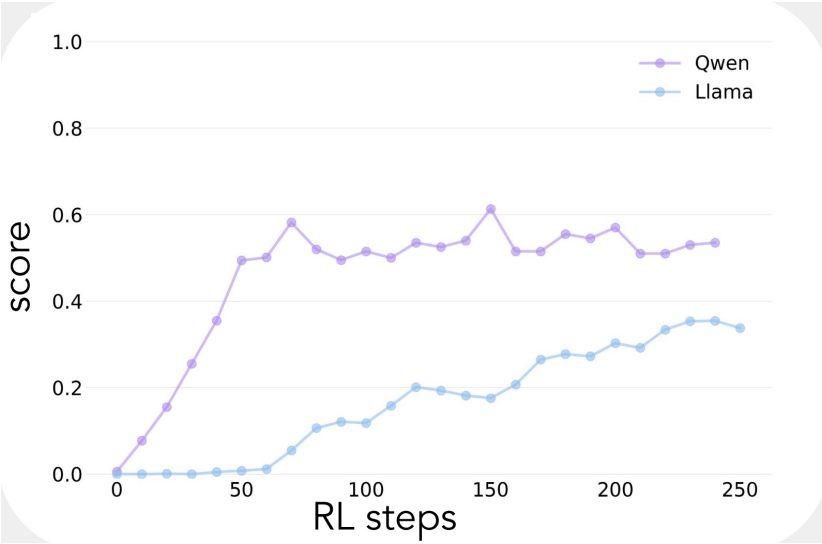
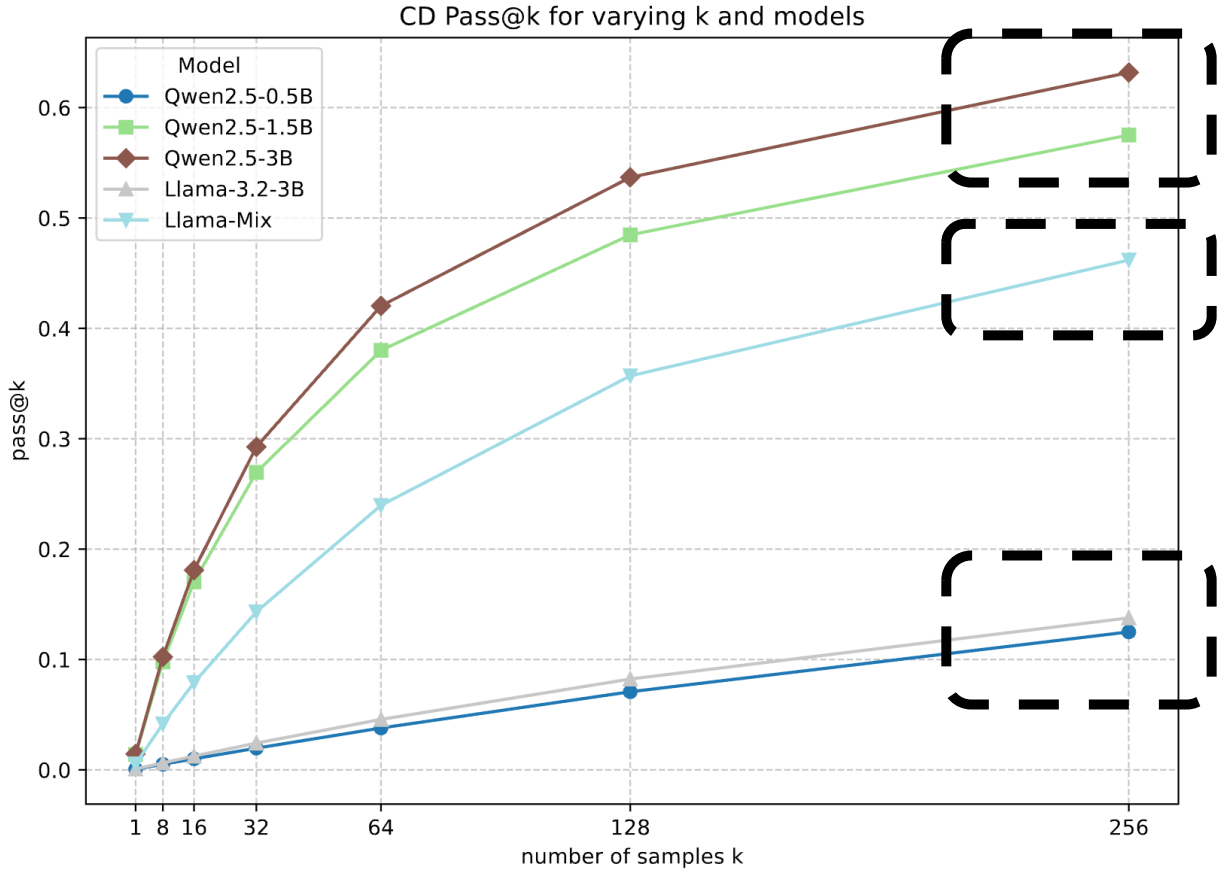


Image from Gandhi et. al

Pass @ k as post-training proxy

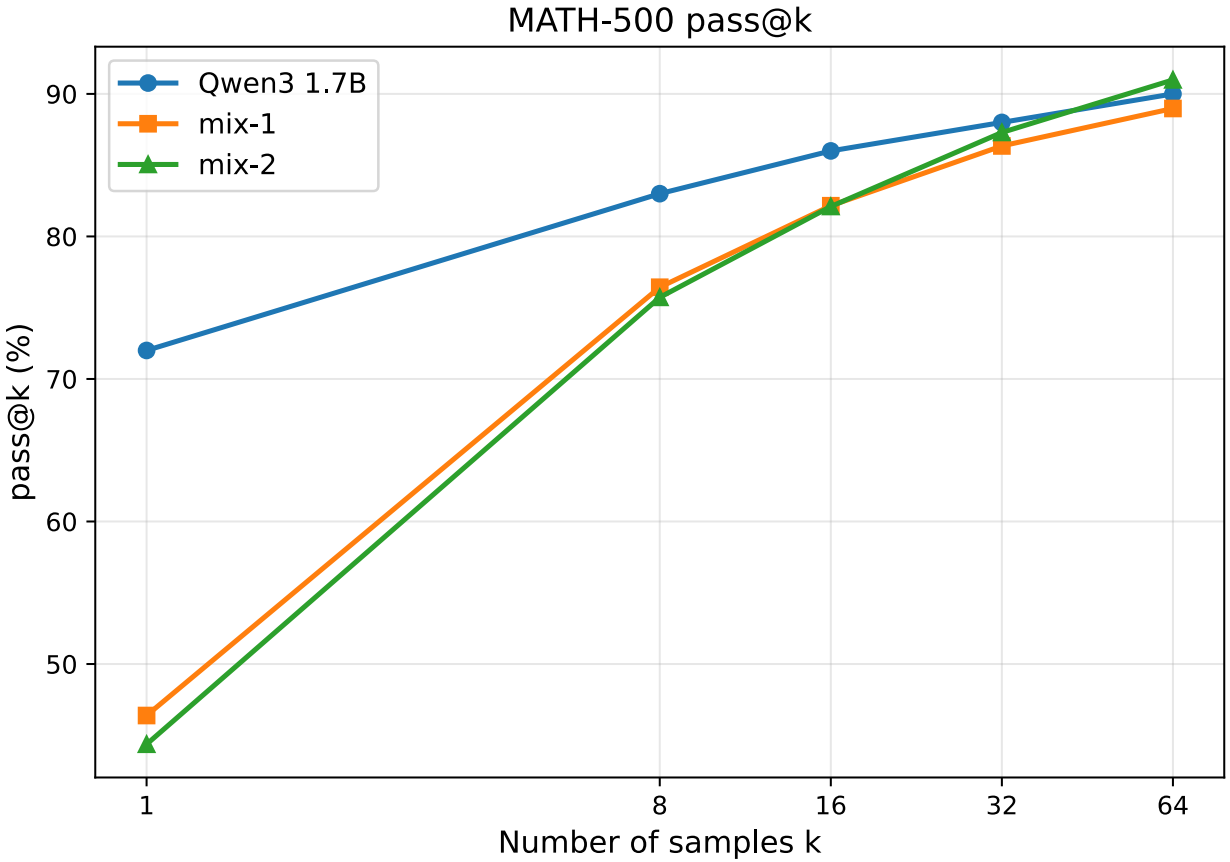


Same as RL high-performers

Early synthetic data CPT tests

Same as RL low-performers

SFT-in-Pre-Training



Recap

- Problem: Choosing meaningful objective for data mix optimization
- Solution 1: Identify pre-training Evals that correlate with final performance
- Solution 2: Train with SFT (chat-formatted) data and measure pass rates

$$\text{Data: } D = \{d_1, \ldots, d_N\}$$

$$\text{Training: } \arg\min_{\theta} \mathbb{E}_{d \sim p_\tau(D)} \ell(\text{LLM}_\theta, d)$$

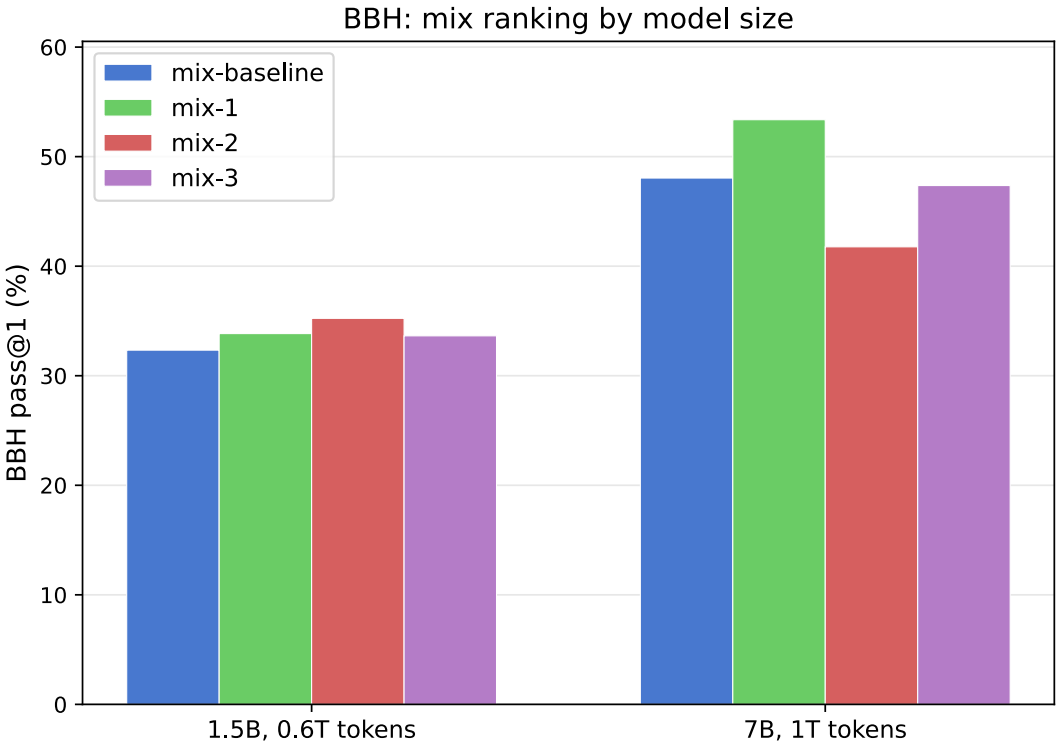
$$p_\tau(D) = \sum_n w_\tau(d_n)\delta_{d_n}$$

$$\text{Data mixing: } \arg\max_{\tau} \text{Evals}(\text{LLM}_\theta) \text{ ✅}$$

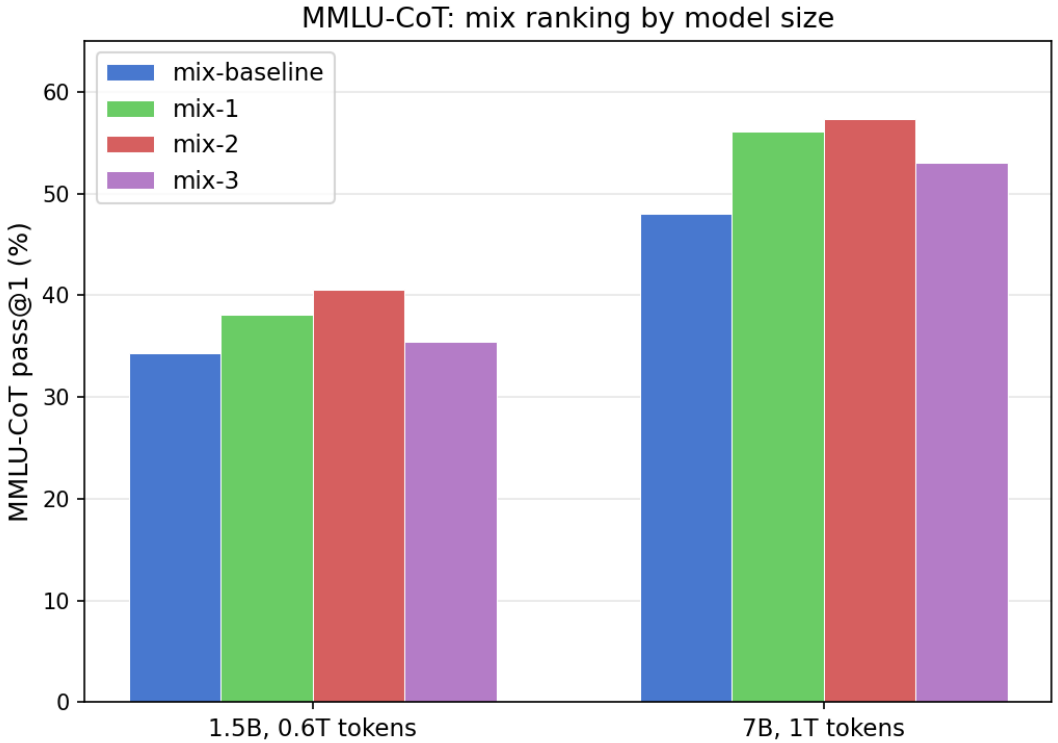
$$\text{such that } \theta = \arg\min_{\theta} \mathbb{E}_{d \sim p_\tau(D)} \ell(\text{LLM}_\theta, d)$$

$$\text{Approach: } f_\eta(\tau) \approx \text{Evals}(\text{LLM}_\theta)$$

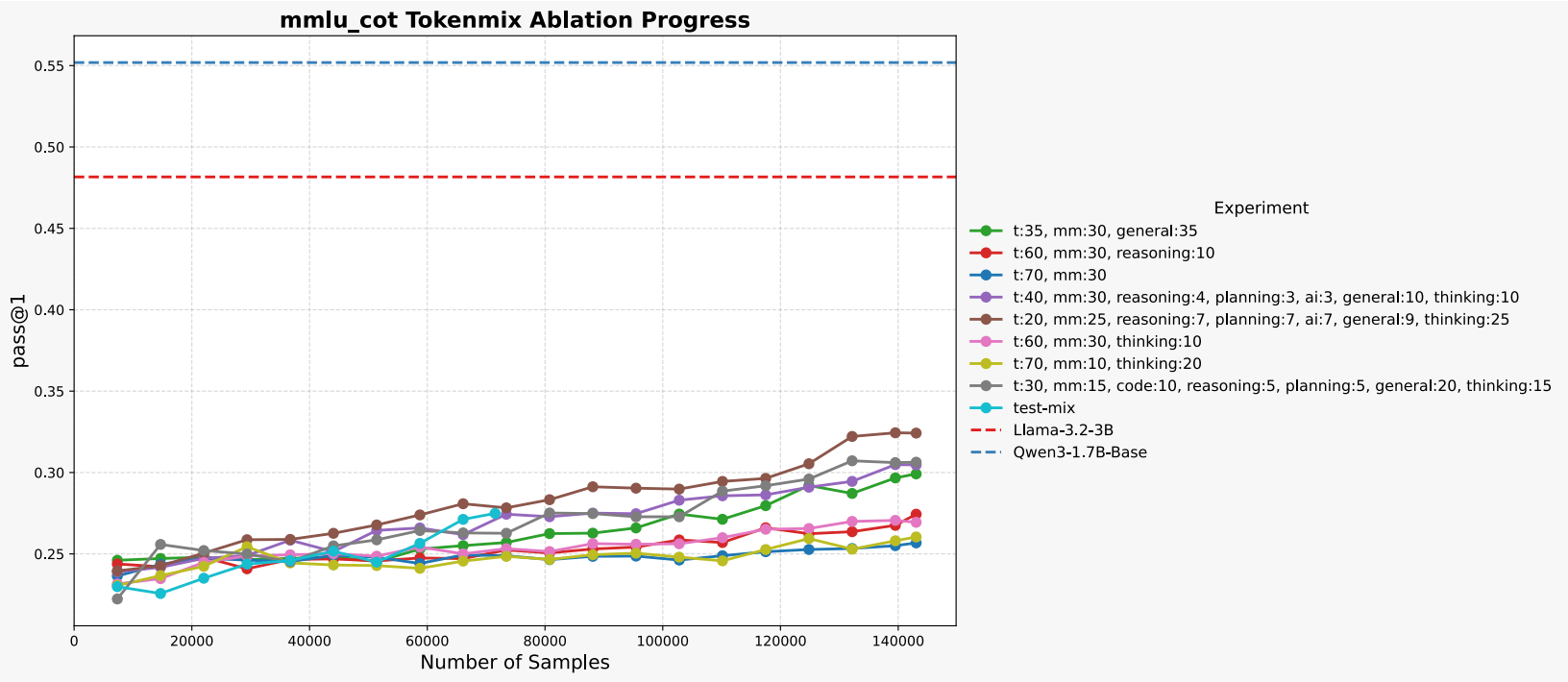
Does optimal Mix depend on data and model size?



Does optimal Mix depend on data and model size?



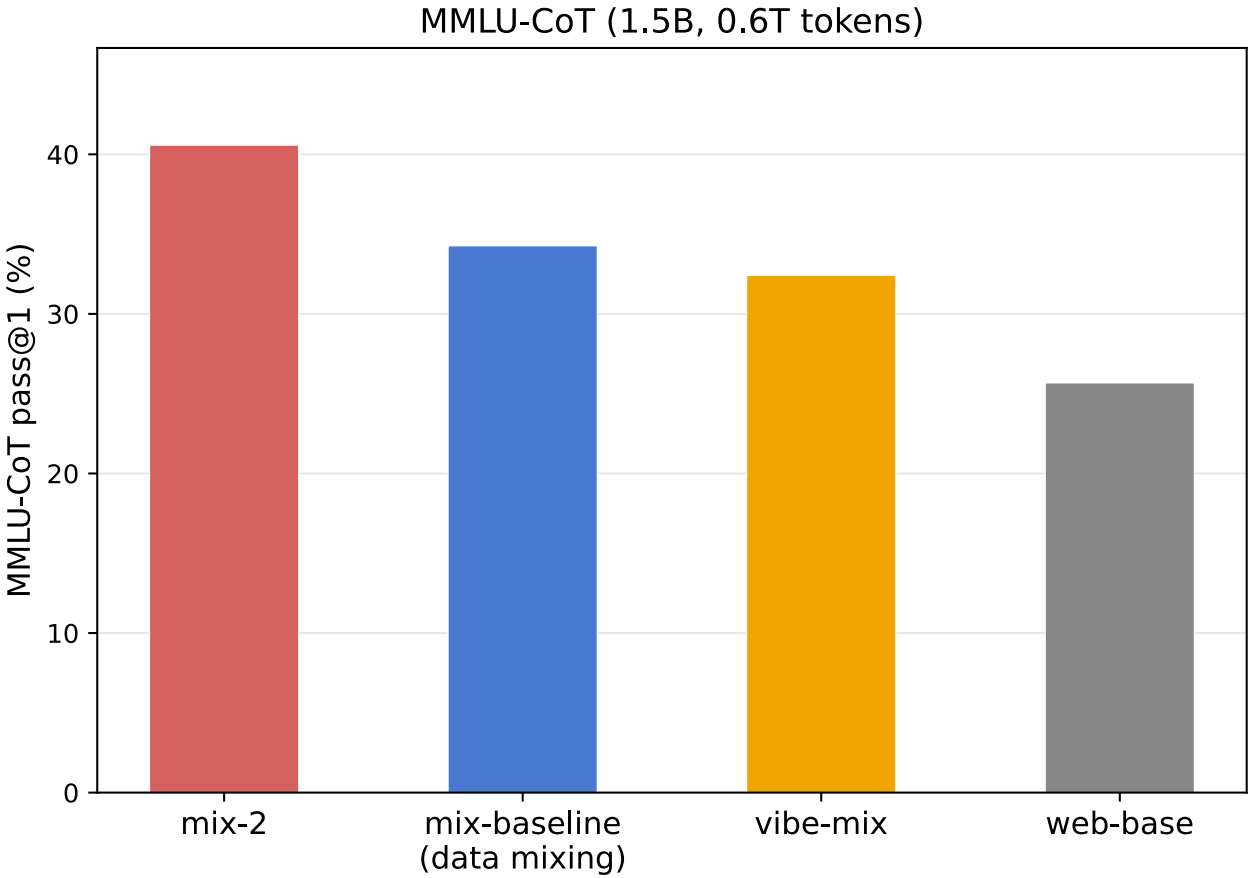
Data Mixing via Vibe-Mixing



Our attempt at Data Mixing

- Fine-grained sources via WebOrganizer
- Constrained sampling to cover feasible set of mixing weights
- Progressive refinement of promising mixing weight candidates
- Evals aggregation accounting for noisy evals, varying difficulty, and eval correlations

It worked but improvement is Small



Data mixing experiment required **375k GPU-hours**

Recap

- Data Mixing: many nuances; computationally expensive; small gains
- Data Diversity: steady and reliable gains

Synthetic Data at Pre-Training Scale

Cascading approaches
(curriculum, ICL, etc): don’t scale well

Key Idea: ground synthetic data in real data

Query-based Synthetic Data

```
    {
        "name": "data_science_problem",
        "instruction": """Come up with a data science problem loosely inspired by concepts related to:

{query}

Create a completely self-contained scenario. Begin by stating a clear data science problem statement.
Then thoroughly describe the available data, study goals, and potential analytical approaches with
justification. Write implementation code with comments. The final output should read as a coherent,
standalone data science case study with no references to the original prompt."""
    },

    {
        'name': 'dual_process_reasoning_system',
        'instruction': """Generate reasoning about the following problem:

{query}

Begin by restating the problem. First provide an intuitive (System 1) assessment, then transition to
a deliberate (System 2) analysis. Show how these two reasoning modes lead to different conclusions
and how they can be reconciled."""
    },
```

How do I have an epic water fight?

Response:

The problem at hand is about how to have an epic water fight. Let's break this down by first using intuitive reasoning, and then transitioning to a more deliberate analysis.

Intuitively (System 1), an epic water fight might seem like a simple idea to execute. This could involve grabbing as many water guns as possible, filling them up at a nearby hose, and then running around with friends or family, spraying each other from various angles [...]

However, adopting a more deliberate (System 2) approach to the problem reveals additional considerations that could elevate a water fight to being truly epic. This involves more structured planning and execution:

1. **Preparation:** Consider the type and quantity of water guns, ensuring there's a mix of short-range and long-range weapons. Additionally, [...]

2. **Location and Setup:** The setting can make a significant difference. A spacious backyard with varied terrain (trees, a pool, or a trampoline) can add to the excitement [...]

3. **Rules and Teams:** [...]

4. **Safety and Supervision:** [...]

5. **Aftermath:** [...]

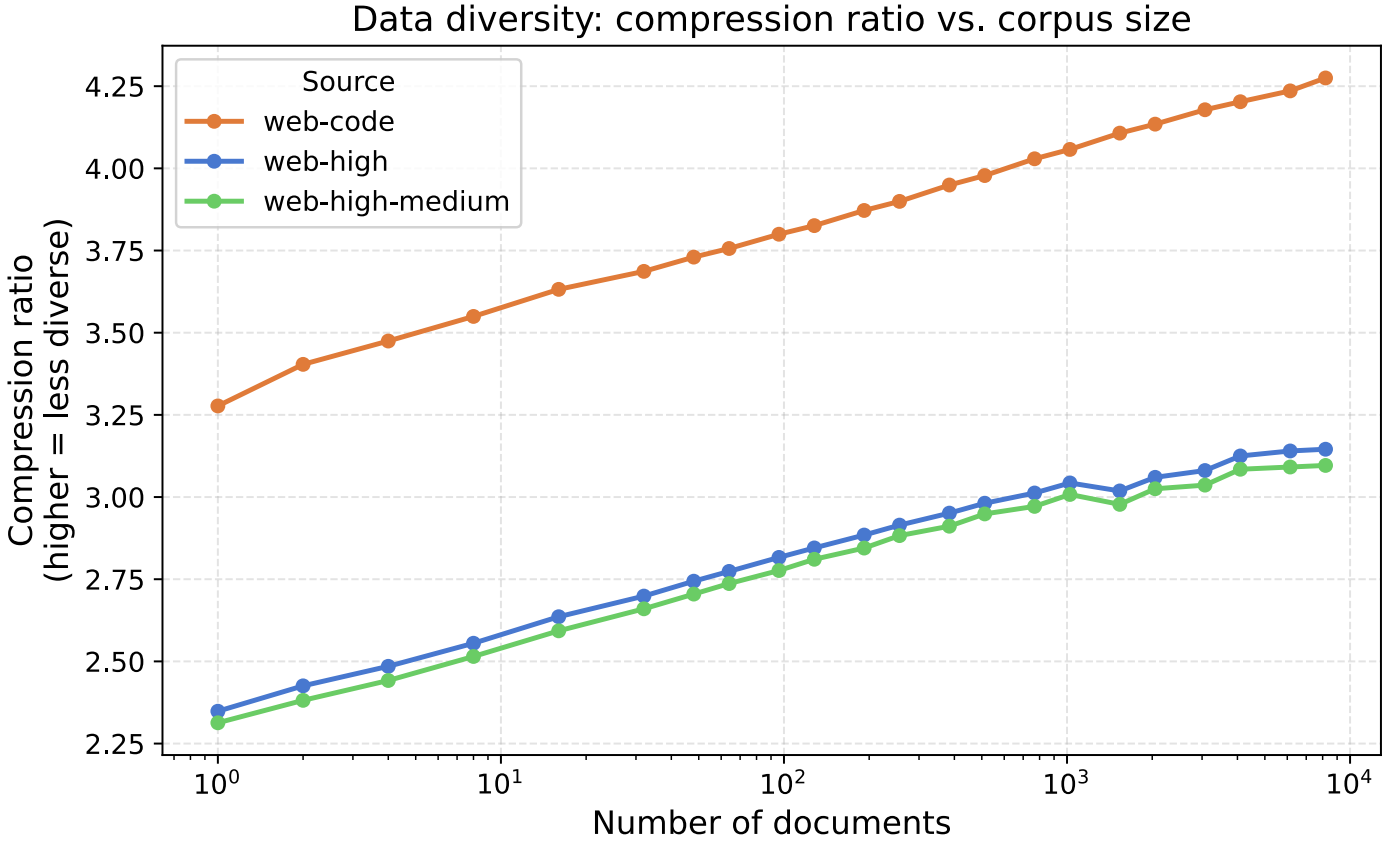
Reconciling the intuitive and deliberate approaches, the key is to balance spontaneity with planning. While the immediate joy and excitement of a water fight are important, thoughtful preparation can enhance the experience [...]

https://huggingface.co/datasets/LLM360/TxT360-Midas

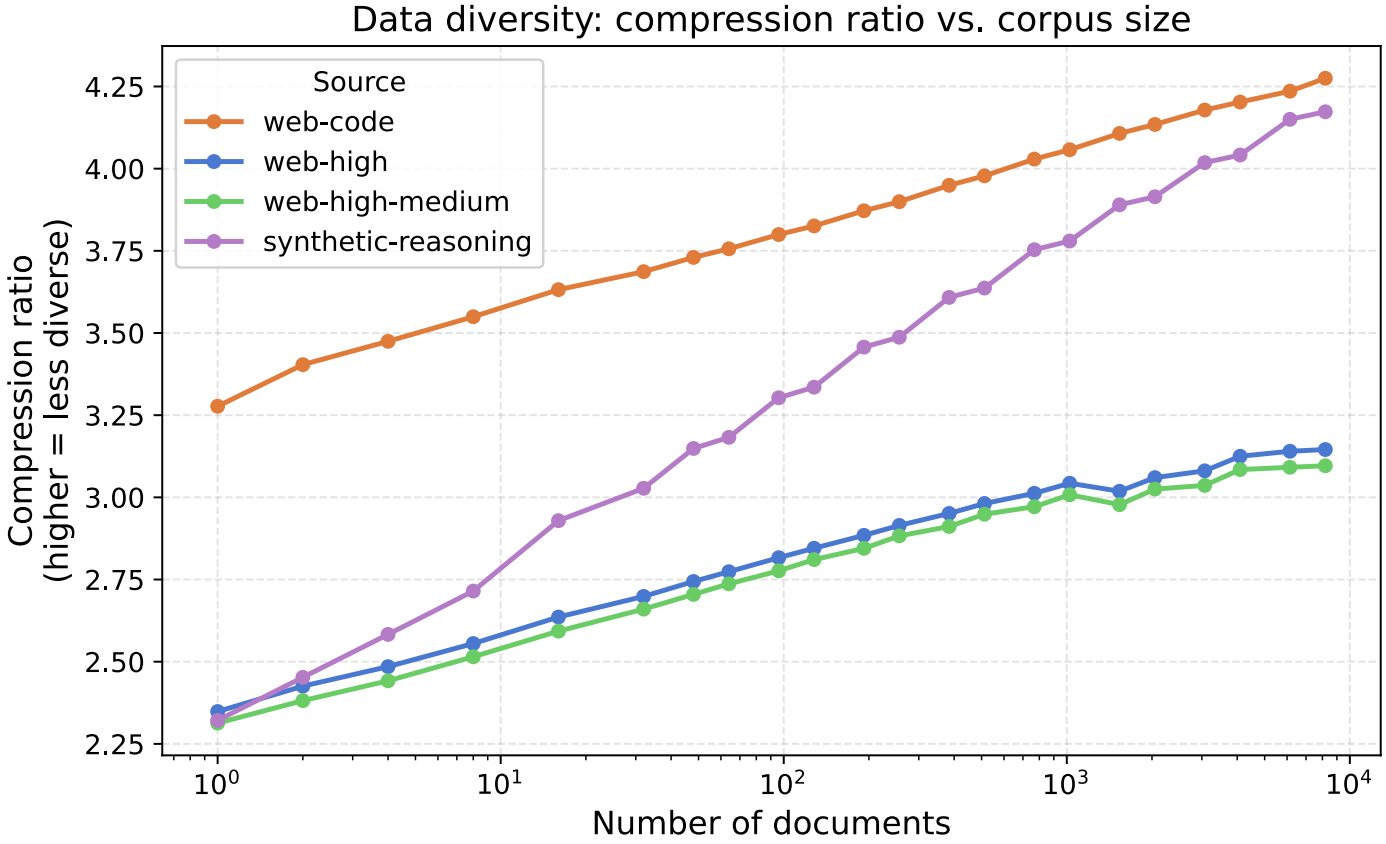
text string
Graph Neural Network Application for Jogging Shoe Recommendation --- ## **1. Executive Summary** This technical document outlines a **Graph Neural Network (GNN)** application designed to recommend jogging shoes tailored to user preferences, health goals…
Ethical AI Assessment Framework for the Chemical Industry: Safety Evaluation of Methyl 6-Amino-5-Bromonicotinate In the evolving landscape of the chemical industry, the integration of artificial intelligence (AI) systems into safety assessments has become…
A/B Testing Framework: Optimizing Blood Type Awareness Campaigns --- ### **1. Business Context** In the healthcare and public health sector, awareness of one's blood type is critical for emergency preparedness, blood donation, and personal health management.…
Let's explore a **generative AI use case** inspired by the query: > **"What is variables in programming? Is it available among various programming languages?"** --- ## ✅ **Use Case Description** ### **Title:** **"Explain Programming Concepts Across Languages…
Federated Learning Proposal: Collaborative Anomaly Detection in Casino Surveillance Systems --- ### **1. Introduction** Casinos operate in high-stakes environments where security is paramount. Traditional surveillance systems rely on centralized video feeds an…
AutoML Framework for Anomalous Human Pattern Detection in a Simulated Alien Environment --- ## **1. Executive Summary** This document outlines a comprehensive **AutoML framework** designed for the **detection and classification of human behavioral patterns…
Automated Machine Learning (AutoML) Framework for Smoke Odor Removal Optimization ## Executive Summary This document presents a comprehensive AutoML framework designed to optimize smoke odor removal strategies in residential environments. The system leverages

< Previous 1 2 3 ... 2,094,315 Next >

Does it Scale?



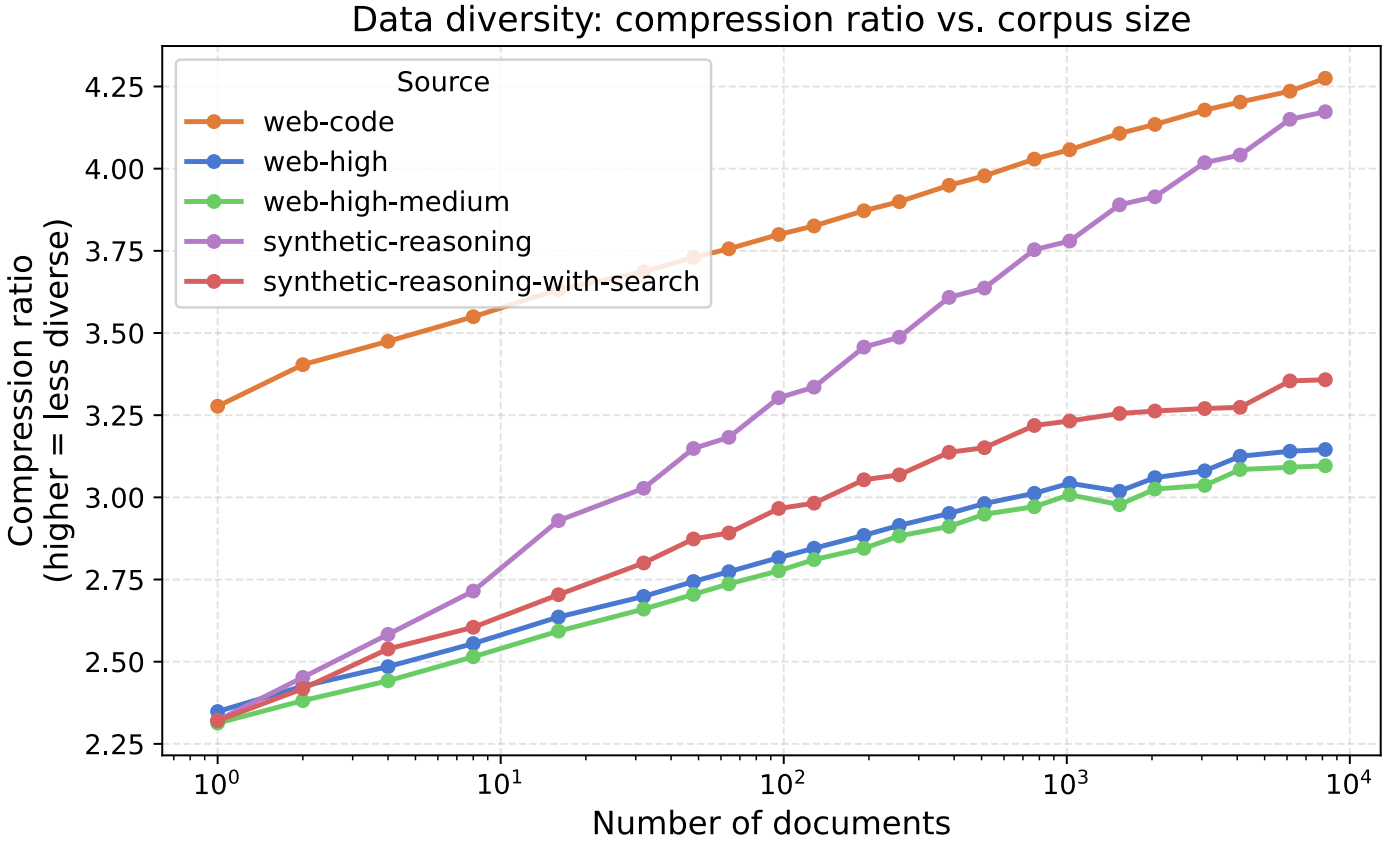
Does it Scale?



More Grounding & Diversity

- Index pre-training data
- Retrieve additional relevant context based on the seed query
- Add randomness to the prompt

Search-Grounding Increases Diversity



If you could train on anything, what would it be?

Popular approach:

- Pre-training: Web
- Mid-training: HQ Web (Books/Wiki)
- SFT: Instruction-Response pairs

If you could train on anything, what would it be?

(Somewhat) Statistical perspective:

- IID generalization is reliable
- OOD generalization is limited

Decide what we want our model to do and go for it!

Thank You!

Questions?

